PA153

Vít Baisa

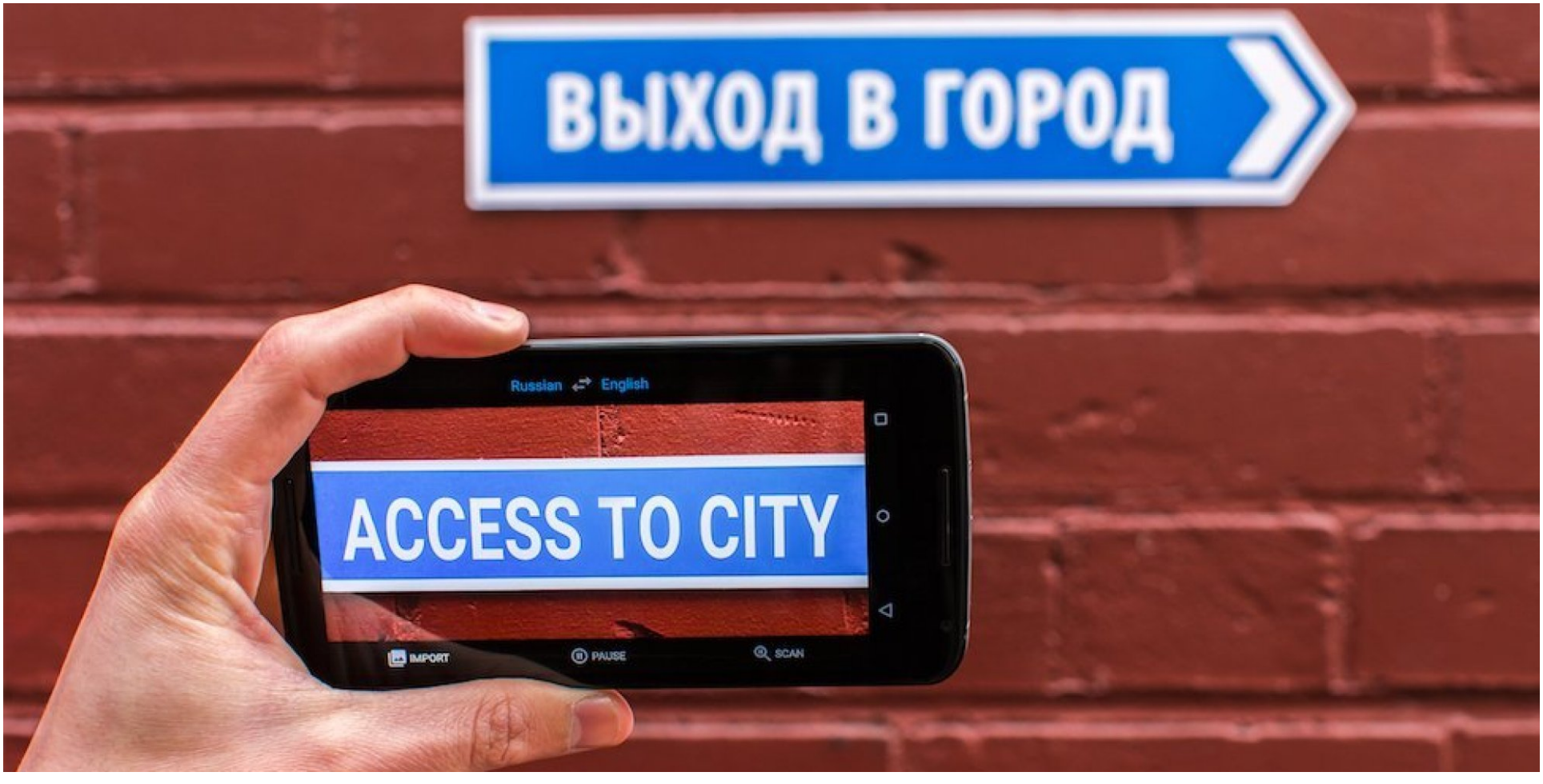
ENGLISH TO CZECH MT

Moses is an implementation of the statistical (or data-driven) approach to machine translation (MT). This is the dominant approach in the field at the moment, and is employed by the online translation systems deployed by the likes of Google and Microsoft.

Mojžíš je implementace statistické (nebo řízené daty) přístupu k strojového překladu (MT). To je převládajícím přístupem v oblasti v současné době, a je zaměstnán pro on-line překladatelských systémů nasazených likes Google a Microsoft.

Moses je implementace statistického (nebo daty řízeného) přístupu k strojovému překladu (MT). V současné době jde o převažující přístup v rámci strojového překladu, který je použit online překladovými systémy nasazenými Googlem a Microsoftem.

Mojžíš je provádění statistické (nebo aktivovaný) přístup na strojový překlad (mt). To je dominantní přístup v oblasti v tuto chvíli, a zaměstnává on - line překlad systémů uskutečněné takové, Google a Microsoft.


ВЫХОД В ГОРОД
Russian
English
ACCESS TO CITY
IMPORT
PAUSE
SCAN

QUESTIONS

- Is accurate translation possible at all?
- What is easier: to translate from / to your mother tongue?
- How we know w_1 is equivalent to w_2?
- English wind types: airstream, breeze, crosswind, dust devil, easterly, gale, gust, headwind, jet stream, mistral, monsoon, prevailing wind, sandstorm, sea breeze, sirocco, southwester, tailwind, tornado, trade wind, turbulence, twister, typhoon, whirlwind, wind, windstorm, zephyr

EXAMPLE OF HARD WORDS

- *alkáč*, *večerníček*, *telka*, *čoklbuřt*, *knížečka*, *ČSSD* ... ?
- *matka, macecha, mamka, máma, maminka, matička, máti, mama, mamča, mamina*
- scvrnkls, nejneobhospo....nějšími
- Navajo Code: language as a cipher
- Leacock: Nonsese novels (Literární poklesky)

MACHINE TRANSLATION

We consider only technical / specialized texts:

- web pages,
- technical manuals,
- scientific documents and papers,
- leaflets and catalogues,
- law texts and
- in general, texts from specific domains.

Nuances on different language levels in art literature are out of scope of current MT systems.

MACHINE TRANSLATION: ISSUES

In fact an output of MT is always revised. We distinguish pre-editing and post-editing.

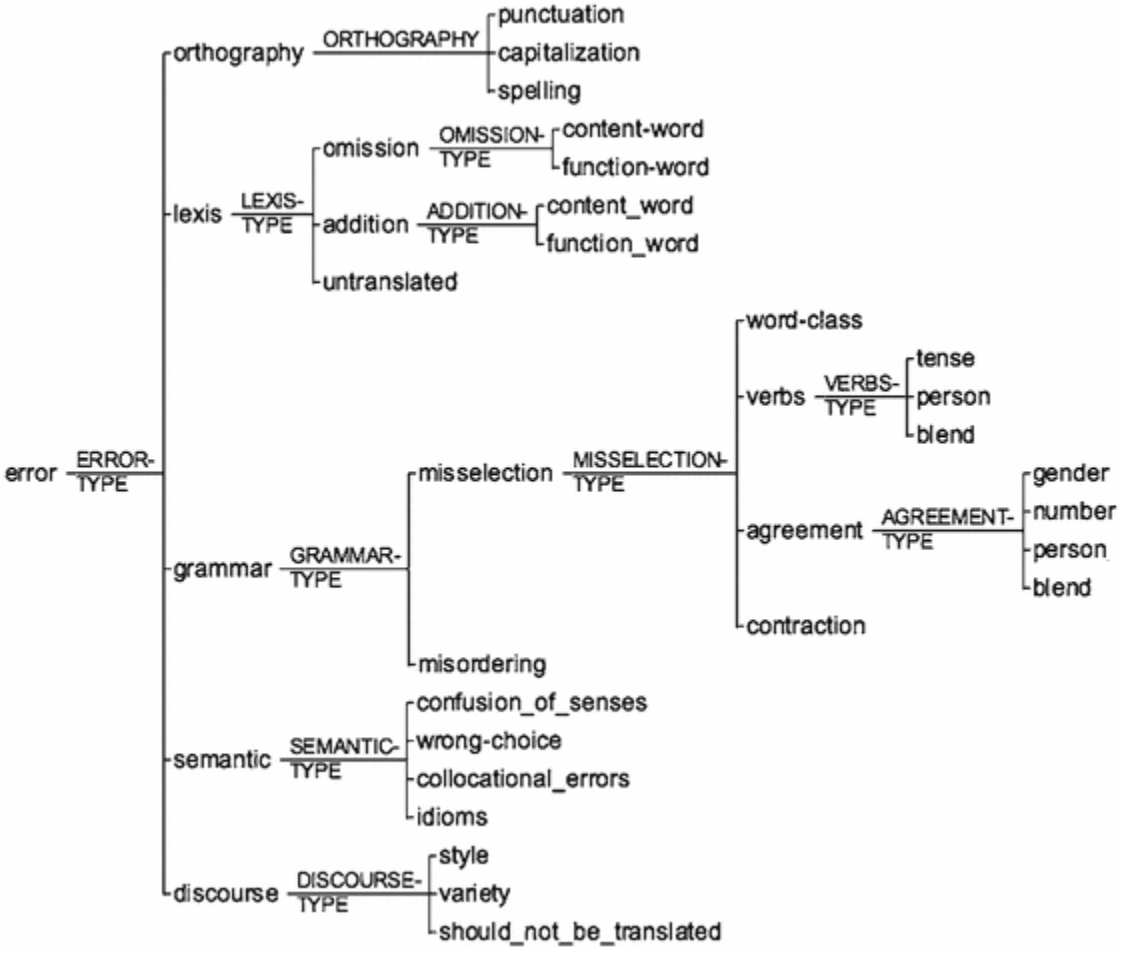
MT systems make different types of errors.

These mistakes are characteristic for human translators:

- wrong prepositions: (*I am in school*)
- missing determiners (*I saw man*)
- wrong tense (*Viděl jsem*: *I was seeing*), ...

For computers, errors in meaning are characteristic:

- *Kiss me honey.* → *Polib mi med.*



Costa, Ângela, et al. "A linguistically motivated taxonomy for Machine Translation error analysis." Machine Translation 29.2 (2015): 127-161.

FREE WORD ORDER

The more morphologically rich language, the freer word order it has.

Katka **snědla** kousek koláče.

- Kati megevett egy szelet tortát → Katie eating a piece of cake
- Egy szelet tortát Kati evett meg → Katie ate a piece of cake
- Kati egy szelet tortát evett meg → Katie ate a piece of cake
- Egy szelet tortát evett meg Kati → Katie ate a piece of cake
- Megevett egy szelet tortát Kati → Katie eating a piece of cake
- Megevett Kati egy szelet tortát → Katie ate a piece of cake

FREE WORD ORDER IN CZECH

- Víš, že se z kávy vyrábí mouka?
- Víš, že se z kávy mouka vyrábí?
- Víš, že se mouka vyrábí z kávy?
- Víš, že se mouka z kávy vyrábí?
- Víš, že se vyrábí mouka z kávy?
- Víš, že se vyrábí z kávy mouka?

How their meanings differ?

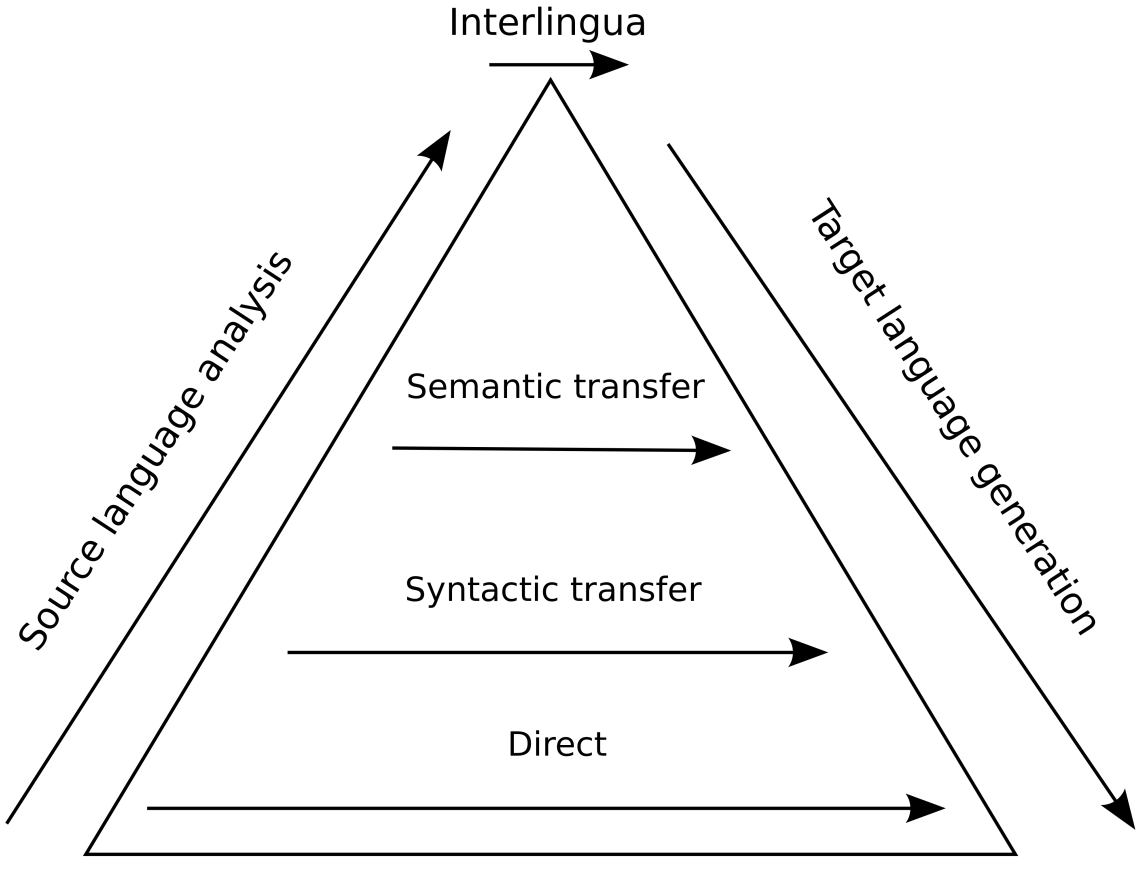
DIRECT METHODS FOR IMPROVING MT QUALITY

- limit input to a:
  - sublanguage (indicative sentences)
  - domain (informatics)
  - document type (patents)
- text pre-processing (e.g. manual syntactic analysis)

CLASSIFICATION BASED ON APPROACH

- rule-based, knowledge-based (RBMT, KBMT)
  - transfer
  - with interlingua
- statistical machine translation (SMT)
- hybrid machine translation (HMT, HyTran)
- neural networks

VAUQUOIS'S TRIANGLE



MACHINE TRANSLATION NOWADAYS

- big companies (Microsoft) focused on English as SL
- large pairs (En:Sp, En:Fr): very good translation quality
- SMT enriched with syntax
- Google Translate as a gold standard
- morphologically rich languages neglected
- En: *a* :En pairs prevail
- neural networks being deployed

MOTIVATION IN 21ST CENTURY

- translation of web pages for gisting (getting the main message)
- methods for speeding-up human translation substantially (translation memories)
- cross-language extraction of facts and search for information
- instant translation of e-communication
- translation on mobile devices

RULE-BASED MT

RULE-BASED MACHINE TRANSLATION

- RBMT
- linguistic knowledge in form of rules
- rules for analysis of SL
- rules for transfer between languages
- rules for generation/rendering/synthesis of TL

KNOWLEDGE-BASED MACHINE TRANSLATION

- systems using linguistic knowledge about languages
- older types, more general notion
- analysis of meaning of SL is crucial
- no total meaning (connotations, common sense)
- to be able to translate *vrána na stromě*
  not necessary to know vrána is a bird and can fly
- term KBMT rather for systems with interlingua
- for us KBMT = RBMT

KBMT CLASSIFICATION

- direct translation
- systems with interlingua
- transfer systems

The only types of MT until 90s.

DIRECT TRANSLATION

- the oldest systems
- one step process: transfer
- Georgetown experiment, METEO
- interest dropped quickly

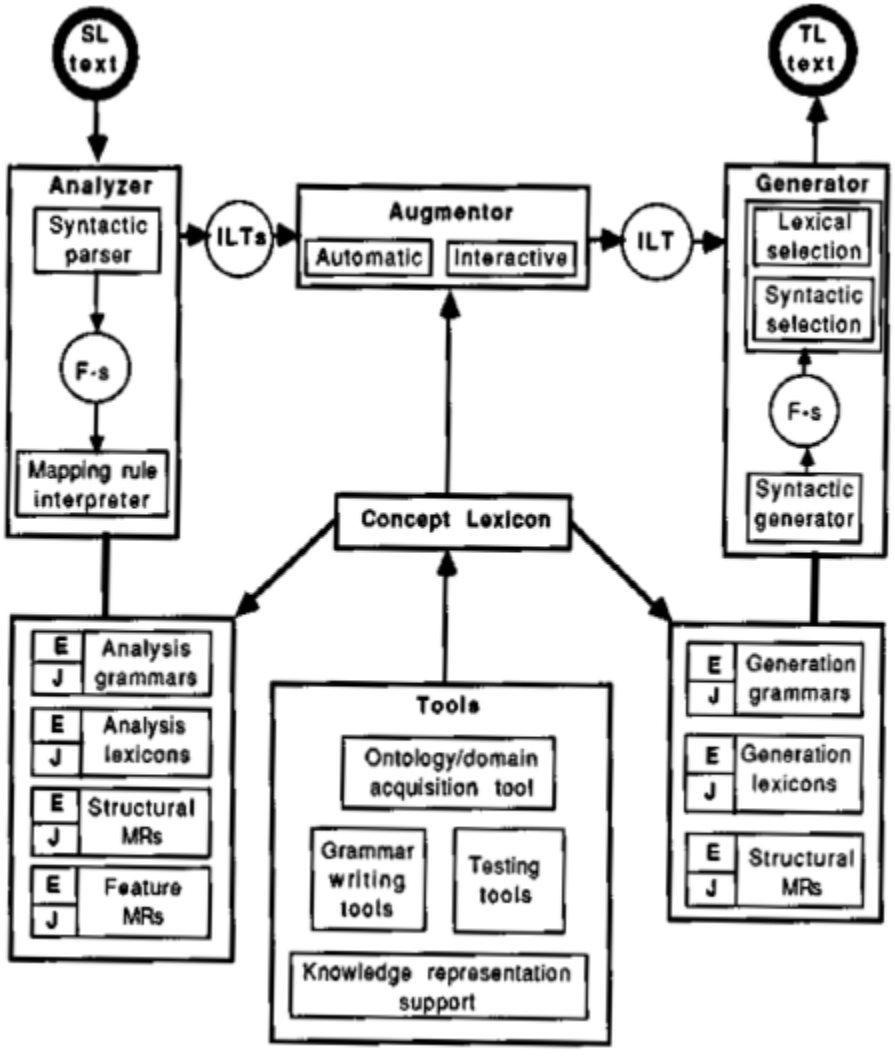
DIRECT TRANSLATION

- focus on S ↔ T elements correspondences
- first experiments on En-Ru pair
- all components are bound to a language pair (and one direction)
- typically consists of:
  - translation dictionary
  - monolithic program dealing with analysis and generation
- necessarily one-directional and bilingual
- efficacy: for N languages we need ?

MT WITH INTERLINGUA

- we suppose it is possible to convert SL to a language-independent representation
- interlingua (IL) must be unambiguous
- two steps: analysis & synthesis (generation)
- from IL, TL is generated
- analysis is SL-dependent but TL-independent
- and vice versa for synthesis
- for translation among N languages

KBMT-89

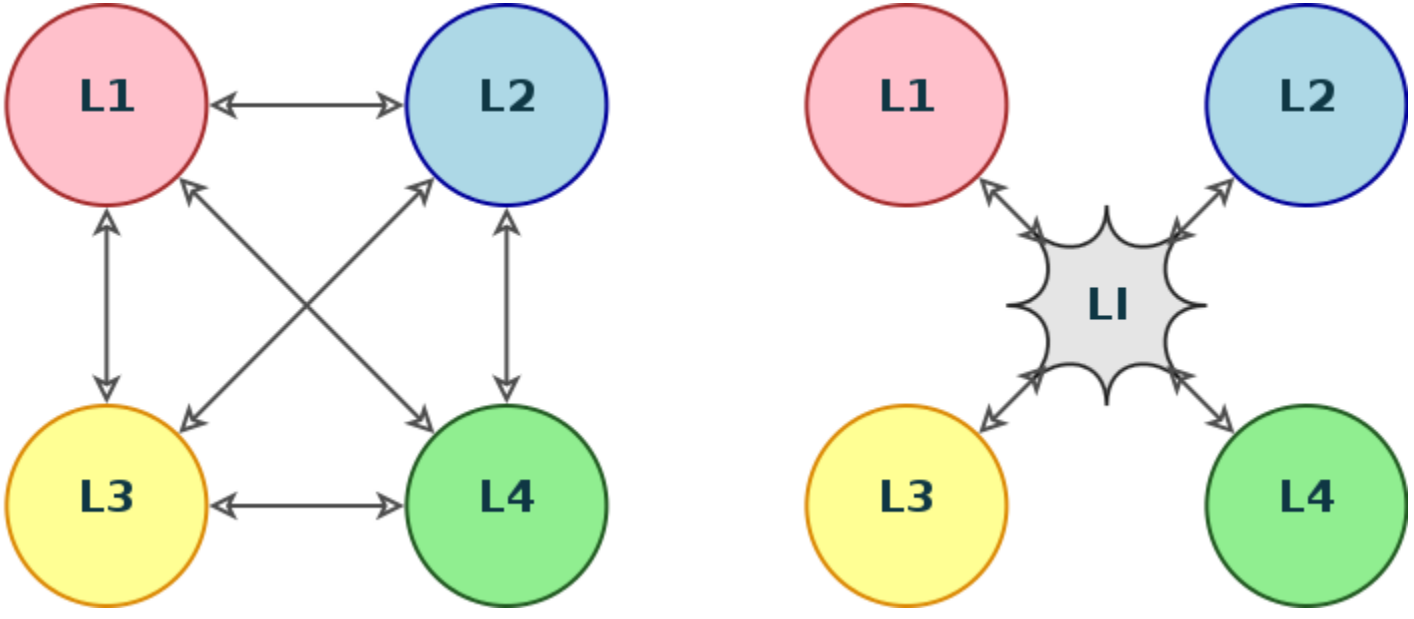


Nirenburg, Sergei. *Knowledge-based machine translation.* Machine Translation 4.1 (1989): 5-24.

TRANSFER TRANSLATION

- analysis up to a certain level
- transfer rules S forms → T forms
- not necessarily between same levels
- usually on syntactic level
  → context constraints
  (not available in direct translation)
- distinction IL vs. transfer blurred
- three-step translation

INTERLINGUA VS. TRANSFER



SOURCE LANGUAGE ANALYSIS

TOKENIZATION

- first level in Vauquois' △
- input text to tokens (words, numbers, punctuation)
- token = sequence of non-white characters
- output = list of tokens
- input for further processing

OBSTACLES OF TOKENIZATION

- don't: do n't, do n 't, don 't, ?
- červeno-černý: červeno - černý, červeno-černý, červeno- černý

SCRIPTIO CONTINUA

คำว่า **ไทย** หมายความว่า อิสรภาพ เสรีภาพ หรืออีกความหมายหนึ่ง คือ ใหญ่ ยิ่งใหญ่ เพราะการจะเป็นอิสระได้จะต้องมีกำลังที่มากกว่า แข็งแกร่งกว่า เพื่อป้องกันการรุกรานจากข้าศึก แม้คำนี้จะมีรูปเหมือนคำยืมจากภาษาบาลีสันสกฤต แต่แท้ที่จริงแล้ว คำนี้เป็นคำไทยแท้ที่เกิดจากกระบวนการสร้างคำที่เรียกว่า 'การลากคำเข้าวัด' ซึ่งเป็นการลากความวิธีหนึ่ง ตามหลักคติชนวิทยา คนไทยเป็นชนชาติที่นับถือกันว่า ภาษาบาลีซึ่งเป็นภาษาที่บันทึกพระธรรมคำสอนของพระพุทธเจ้าเป็นภาษาอันศักดิ์สิทธิ์และเป็นมงคล เมื่อคนไทยต้องการตั้งชื่อประเทศว่า **ไท** ซึ่งเป็นคำไทยแท้ จึงเติมตัว **ย** เข้าไปข้างท้าย เพื่อให้มีลักษณะคล้ายคำในภาษาบาลีสันสกฤตเพื่อความเป็นมงคลตามความเชื่อของตน ภาษาไทยจึงหมายถึงภาษาของชนชาติไทยผู้เป็นไทนั่นเอง

What is word?

TOKENIZATION

- in most cases a heuristic is used
- alphabetic writing systems: split on spaces and on other punctuation marks ?!.,-()/:;
- demo: unitok.py

SENTENCE SEGMENTATION

- MT almost always uses sentences
- 90% of periods are sentence boundary indicators (Riley 1989)
- using list of punctuation (!?.<>)
  Měl jsem 5 (sic!) poznámek.
- exceptions:
  - abbreviations (aj. atd. etc. e.g.)
  - degrees (RNDr., prof.)
- HTML elements might be used (p, div, td, li)
- demo: tag_sentences
- paper on tokenization

OBSTACLES OF SENTENCE SEGMENTATION

- Zeleninu jako rajče, mrkev atd. Petr nemá rád.
- Složil zkoušku a získal titul Mgr. Petr mu dost záviděl.
- John F. Kennedy = one token?
  John F. Kennedy's
- related to named entity recognition
- neglected step in the processing (DCEP, EUR-Lex)

MORPHOLOGICAL LEVEL

MORPHOLOGY

- morpheme: *the smallest item carrying a meaning*
- pří-lež-it-ost-n-ým-i
- prefix-root-infix-suffix-suffix-suffix-affix
- case, number, gender, lemma, affix

MORPHOLOGIC LEVEL

- second level in Vauquois' △
- reducing the immense amount of wordforms
- demo: lexicon sizes of various corpora
- conversion from wordforms to lemmata
  give, gives, gave, given, giving → give
  dělá, dělám, dělal, dělaje, dělejme, ... → dělat
- analysis of grammatical categories of wordforms
  dělali → dělat + past t. + continuous + plural + 3rd p.
  did → do + past t. + perfective + person ? + number ?
  Robertovým → Robert + case ? + adjective + number ?
- demo: wwwajka

MORPHOLOGIC ANALYSIS

- for each token we get a base form, grammar categories, segmentation to morphemes
- What is a base form? Lemma.
- nouns: singular, nominative, positive, masculine
- bycha → bych?, nejpomalejšími → pomalý neschopný → schopný?, mimochodem → mimochod?
- verbs: infinitive
- neraď → radit?, bojím se → bát (se)
- Why infinitive? the most frequent form of verbs
- example

MORPHOLOGICAL TAGS, TAGSET

- language-dependent (various morphological categories)
- attribute system: pairs category--value
  maminkou k1gFnSc7
  udělány k5eAaPmNgFnP
- positional system: 16 fixed positions
  kontury NNFP1-----A----
  zdají VB-P---3P-AA---
- Penn Treebank tagset (English): limited set of tags
  faster RBR
  doing VBG
- CLAWS tagset (English)
- and others (German)
  gigantische ADJA.ADJA.Pos.Acc.Sg.Fem
  erreicht VVPP.VPP.Full.Psp

MORPHOLOGICAL POLYSEMY

- in many cases: words have more than one tag
- PoS polysemy (>1 lemma), in Czech
  jednou k4gFnSc7, k6eAd1, k9
  ženu k1gFnSc4, k5eAaImIp1nS
  k1 + k2, k3 + k5?
- what about English?
- demo: SkELL auto PoS
- polysemy within a PoS
- in Czech: nominative = accusative
  víno: k1gNnSc1, k1gNnSc4, ...
  odhalení: 10 tags

MORPHOLOGICAL DISAMBIGUATION

- for each word: one tag and one lemma
- morphological disambiguation
- a tool: *tagger*
- translational polysemy is another issue
  pubblico – Öffentlichkeit, Publikum, Zuschauer
- most of methods use *context*
- i.e. surrounding words, lemmas, tags

STATISTICAL DISAMBIGUATION

- the most probable sequence of tags
  Ženu je domů.
  k5|k1, k3|k5, k6|k1
  Mladé muže
  gF|gM, nS|nP
- there are tough situations: dítě škádlí lvíče
- machine learning trained on manually tagged/disambiguated data
- Brill's tagger, TreeTagger, Freeling, RFTagger
- demo for Czech: Desamb (hybrid, DESAM)

RULE-BASED DISAMBIGUATION

- the only option if an annotated corpus not available
- also used as a filter before a statistical method
- rules help to capture wider context
- case, number and gender agreement in noun phrases
  malému (c3, gIMN) chlapci (nPc157, nSc36, gM)
- a more matured: valency structure of sentences
  valency: vidět koho/co, to give OBJ to DIROBJ
  vidím dům → c4
  I gave the present to her → DIROBJ
- VerbaLex, PDEV

GUESSER

- we aim at high coverage: as many words as possible
- for out-of-vocabulary (OOV) tokens
- new, borrowed, compound words
- stemming, guessing PoSes from word suffix
- *vygooglit*, *olajkovat*, *zaxzovat*
- sedm *dunhillek*
- *třitisícedvěstědevadesátpět* znaků
- funny errors: Matka božit, topit box

MORPHOLOGICAL DISAMBIGUATION—EXAMPLE

slovo	analýzy	disambiguace
Pravidelné	k2eAgMnPc4d1, k2eAgInPc1d1, k2eAgInPc4d1, k2eAgInPc5d1, k2eAgFnSc2d1, k2eAgFnSc3d1, k2eAgFnSc6d1, k2eAgFnPc1d1, k2eAgFnPc4d1, k2eAgFnPc5d1, k2eAgNnSc1d1, k2eAgNnSc4d1, k2eAgNnSc5d1, ... (+ 5)	k2eAgNnSc1d1
krmení	k2eAgMnPc1d1, k2eAgMnPc5d1, k1gNnSc1, k1gNnSc4, k1gNnSc5, k1gNnSc6, k1gNnSc3, k1gNnSc2, k1gNnPc2, k1gNnPc1, k1gNnPc4, k1gNnPc5	k1gNnSc1
je	k5eAaImIp3nS, k3p3gMnPc4, k3p3gInPc4, k3p3gNnSc4, k3p3gNnPc4, k3p3gFnPc4, k0	k5eAaImIp3nS
pro	k7c4	k7c4
správný	k2eAgMnSc1d1, k2eAgMnSc5d1, k2eAgInSc1d1, k2eAgInSc4d1, k2eAgInSc5d1, ... (+ 18)	k2eAgInSc4d1
růst	k5eAaImF, k1gInSc1, k1gInSc4	k1gInSc4
důležité	k2eAgMnPc4d1, k2eAgInPc1d1, k2eAgInPc4d1, k2eAgInPc5d1, k2eAgFnSc2d1, k2eAgFnSc3d1, k2eAgFnSc6d1, k2eAgFnPc1d1, k2eAgFnPc4d1, k2eAgFnPc5d1, k2eAgNnSc1d1, k2eAgNnSc4d1, k2eAgNnSc5d1, ... (+ 5)	k2eAgNnSc1d1

PROBLEMS WITH POSES

- quality of MA affects all further levels of analysis
- quality depends on a language (English vs. Hungarian)
- chončaam: my small house (Tajik)
- kahramoni: you are hero (Tajik)
- legeslegmagasabb: the very highest (Hungarian)
- raněný: SUBS / ADJ
- the big red fire truck: SUBS / ADJ?
- The Duchess was entertaining last night.
- Pokojem se neslo tiché pšššš.

MORPHOLOGY—SUMMARY

- MA introduce critical errors into the analysis
- the goal is to limit the immense amount of wordforms
- wordform → lemma + tag
- much simpler for English (cc. 35 tags)
- PoS tagging accuracy depends on a language
- usually around 95%

LEXICAL LEVEL

DICTIONARIES

DICTIONARIES IN MT I

- connection between languages
- transfer systems: syntactic level
- dictionaries crucial for KBMT systems
- GNU-FDL slovník
- Wiktionary

DICTIONARIES IN MT II

- how many items in a dict do we need / want?
  → named entities, slang, MWE
- *listeme*: lexical item, which can not be deduced from the principle of compositionality (slaměný vdovec)
- which form in a dict? → lemmatization
- how many different senses is reasonable to distinguish? → granularity

POLYSEMY IN DICTIONARIES

- words relates to senses
- what is meaning of meaning?
- we need a formal definition for computers
- data is discrete, meaning is continuous
- *man*: an adult male person
  what about 17-years-old male person?

SMOOTH SENSE TRANSITIONS

log

log chair

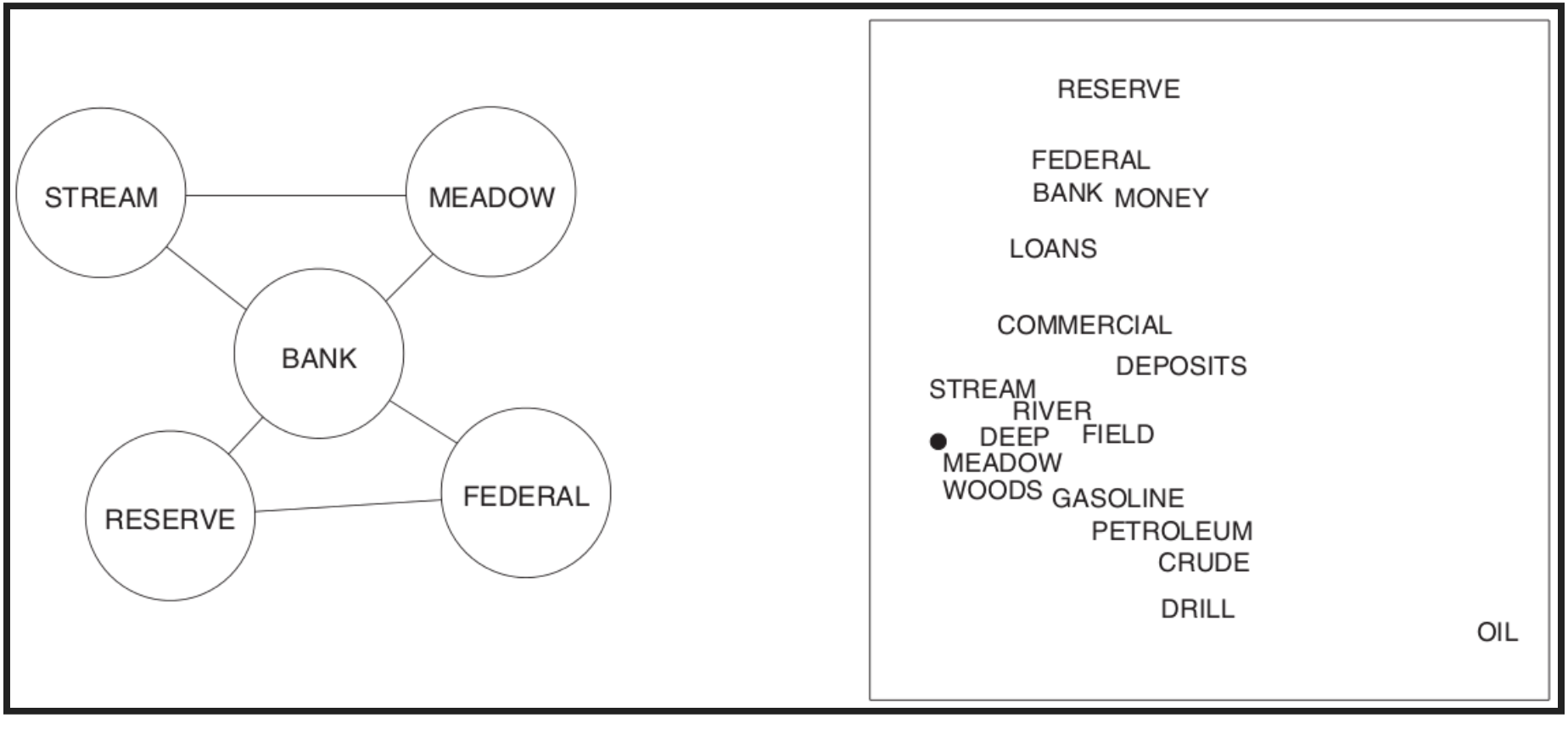
chair

POLYSEMY ON SEVERAL LEVELS

- morphology: *-s*
- word level: *key*
- multiword expressions: *bílá vrána*
- sentence level: *I saw a man with a telescope.*
- homonymy: accidental
  - full homonymy: *líčit*, *kolej*
  - partial homonymy: *los*, *stát*
- polysemy is natural and ubiquitous

MEANING REPRESENTATION

- list: a common dictionary
- graph: senses:vertices, semantic relations:edges
- space: senses:dots, similarity:distance



WORD SENSE DISAMBIGUATION

- finding a proper sense of a word in a given context
- trivial for human, very hard for computers
- we need a finite inventory of senses
- accuracy about 90%
- crucial task for KBMT:
  Ludvig dodávka Beethoven, kiss me honey, …
  box in the pen (Bar-Hillel)
- granularity affects the quality of WSD

SYNTACTIC LEVEL

Miloš and Vojta

SEMANTIC LEVEL / ANALYSIS

Zuzka Nevěřilová, Adam Rambousek

TECTOMT

- PDT formalism, high modularity
- splitting tasks to a sequence of blocks—scenarios
- blocks are Perl scripts communicating via API
- blocks allow massive data processing, parallelisation
- rule-based, statistical, hybrid methods
- processing: conversion to the tmt format → application of a scenario → conversion to an output format

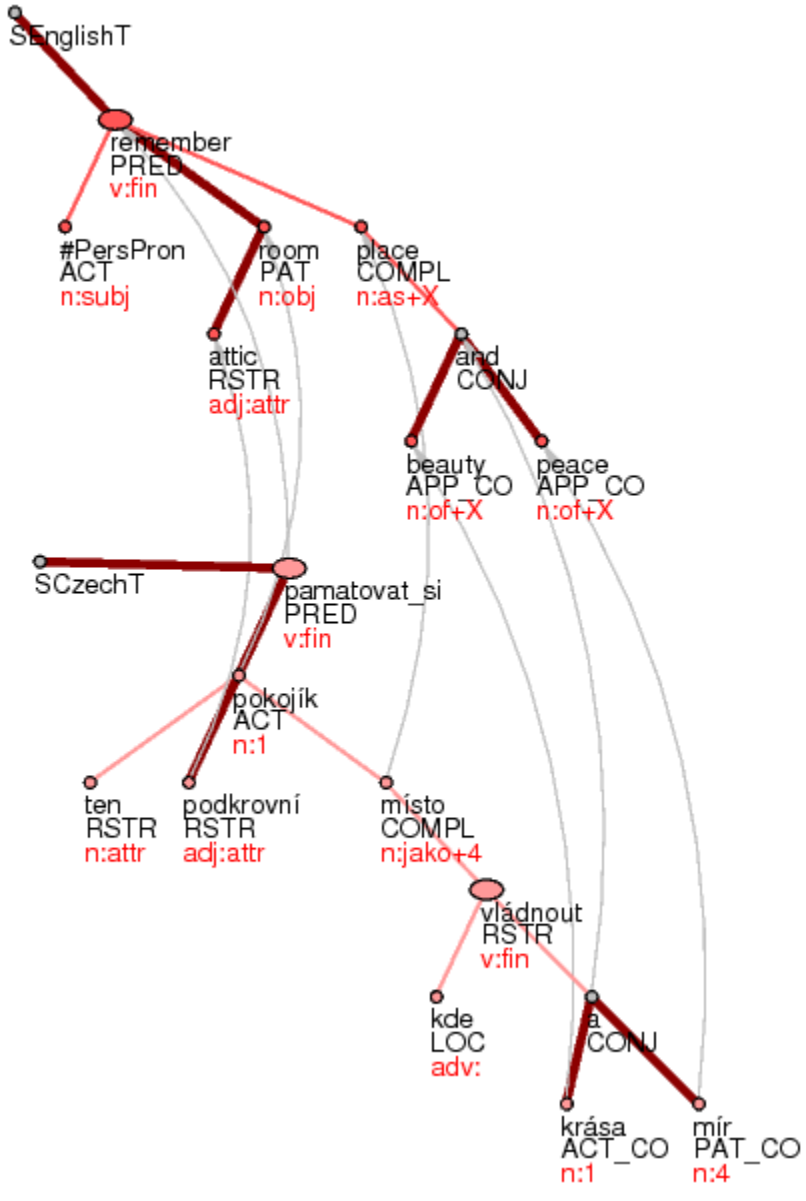
TECTOMT: A SIMPLE BLOCK

English negative particles → verb attributes

```
sub process_document {
  my ($self,$document) = @_;

  foreach my $bundle ($document->get_bundles()) {
    my $a_root = $bundle->get_tree('SEnglishA');

    foreach my $a_node ($a_root->get_descendants) {
      my ($eff_parent) = $a_node->get_eff_parents;
      if ($a_node->get_attr('m/lemma')=~/^(not|n't)$/
          and $eff_parent->get_attr('m/tag')=~/^V/ ) {
        $a_node->set_attr('is_aux_to_parent',1);
      }
    }
  }
}
```


SEnglishT
remember
PRED
v:fin
#PersPron
ACT
n:subj
room
PAT
n:obj
place
COMPL
n:as+X
attic
RSTR
adj:attr
and
CONJ
beauty
APP_CO
n:of+X
peace
APP_CO
n:of+X
SCzechT
pamatovat_si
PRED
v:fin
pokojík
ACT
n:1
ten
RSTR
n:attr
podkrovní
RSTR
adj:attr
místo
COMPL
n:jako+4
vládnout
RSTR
v:fin
kde
LOC
adv:
a
CONJ
krása
ACT_CO
n:1
mír
PAT_CO
n:4

RULE-BASED SYSTEMS: CONCLUSION

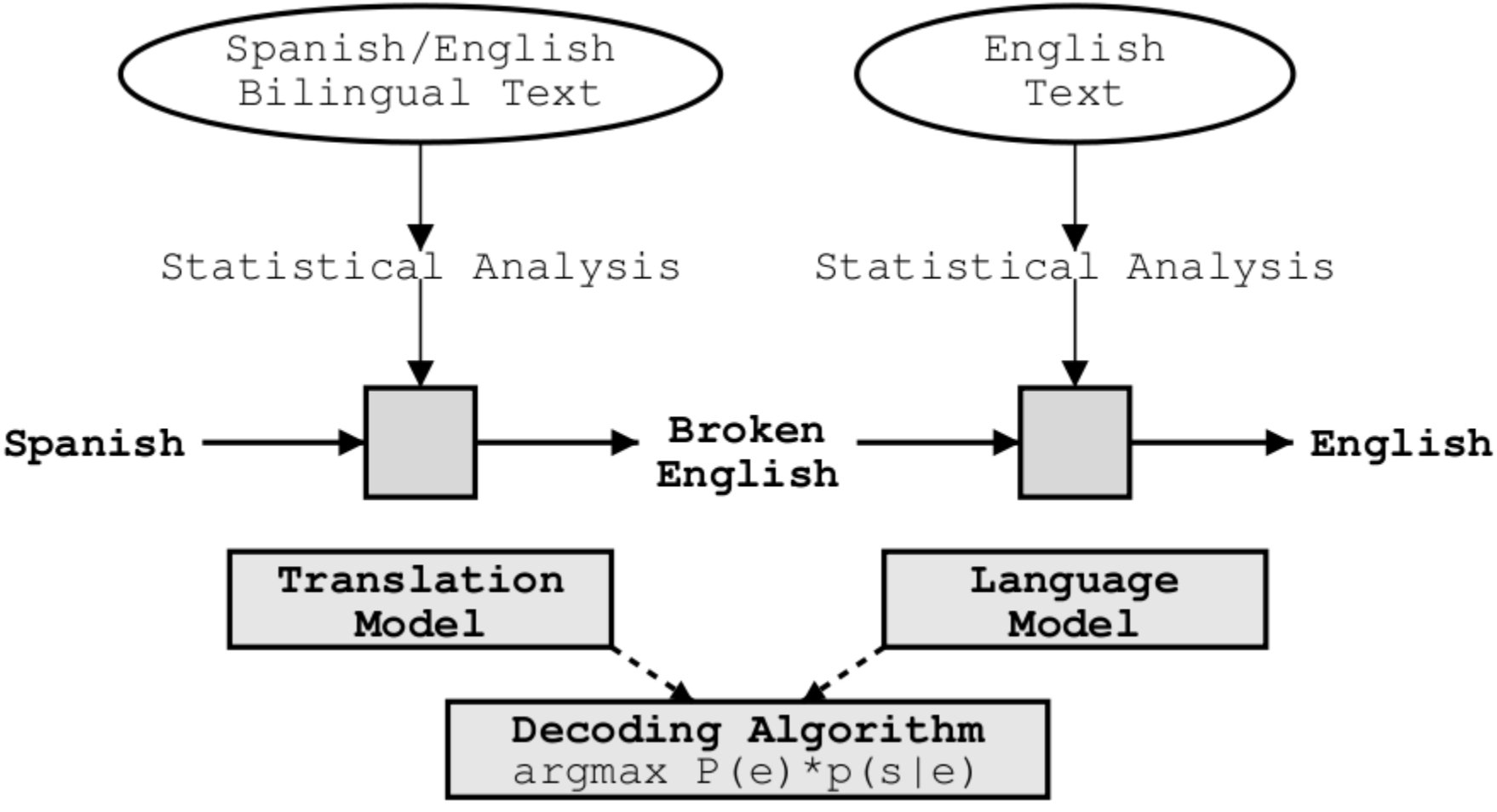
- (purely) rule-based systems not used anymore
- statistical systems achieve better results
- still, some methods from RBMT may improve SMT

STATISTICAL MACHINE TRANSLATION

INTRODUCTION

- rule-based systems motivated by linguistic theories
- SMT inspired by information theory and statistics
- Google, IBM, Microsoft develop SMT systems
- millions of webpages translated with SMT daily
- gisting: we don't need exact translation, sometimes a gist of a text is enough (on of the most frequent use of SMT)
- SMT in assisted MT (CAT)
- trending right now: neural network models for MT
- data-driven approach more viable than RBMT

SMT SCHEME


Spanish/English
Bilingual Text
English
Text
Statistical Analysis
Statistical Analysis
Spanish
Broken
English
English
Translation
Model
Language
Model
Decoding Algorithm
argmax P(e)*p(s|e)

PARALLEL CORPORA I

- basic data source for SMT
- available sources ~10–100 M
- size depends heavily on a language pair
- multilingual webpages (online newspapers)
- paragraph and sentence alignment needed

PARALLEL CORPORA II

- Europarl: 11 ls, 40 M words
- OPUS: parallel texts of various origin, open subtitles, UI localizations
- Acquis Communautaire: law documents of EU (20 ls)
- Hansards: 1.3 M pairs of text chunks from the official records of the Canadian Parliament
- EUR-Lex
- comparable corpora...

SENTENCE ALIGNMENT

- sometimes sentences are not in 1:1 ratio in corpora
- Church-Gale alignment
- hunalign

P	alignment
0.89	1:1
0.0099	1:0, 0:1
0.089	2:1, 1:2
0.011	2:2

SMT NOISY CHANNEL PRINCIPLE

Claude Shannon (1948), self-correcting codes transfered through noisy channels based on information about the original data and errors made in the channels.

Used for MT, ASR, OCR. Optical Character Recognition is erroneous but we can estimate what was damaged in a text (with a language model); errors l↔1↔I, rn↔m etc.

$$
\begin{aligned}
e^* &= \arg\max_e p(e|f) \\
&= \arg\max_e \frac{p(e)p(f|e)}{p(f)} \\
&= \arg\max_e p(e)p(f|e)
\end{aligned}
$$

SMT COMPONENTS I

- language model
- how we get $p(e)$ for any string e
- the more e looks like proper language the higher $p(e)$ should be
- issue: what is $p(e)$ for an unseen e?

SMT COMPONENTS II

- translation model
- for e and f compute $p(f|e)$
- the more e looks like a proper translation of f, the higher $p(f|e)$

SMT COMPONENTS III

- decoding algorithm
- based on TM and LM, find a sentence f as the best translation of e
- as fast as possible and with as few memory as possible
- prune non-perspective hypotheses
- but do not lost any valid translations

LANGUAGE MODELS

WHAT IT IS GOOD FOR?

What is the probability of utterance of $\boldsymbol{s}$?

I go to home vs. *I go home*

What is the next, most probable word?

Ke snídani jsem měl celozrnný …

{ chléb > pečivo > zákusek > mléko > babičku}

CHOMSKY WAS WRONG

Colorless green ideas sleep furiously
vs. *Furiously sleep ideas green colorless*

LM assigns higher p to the 1st! (Mikolov, 2012)

GENERATING RANDOM TEXT

To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have Every enter now severally so, let. (unigrams)

Sweet prince, Falstaff shall die. Harry of Monmouth's grave. This shall forbid it should be branded, if renown made it empty. (trigrams)

Can you guess the author of the original text?

CBLM

MAXIMUM LIKELIHOOD ESTIMATION

$$p(w_3|w_1, w_2) = \frac{count(w_1, w_2, w_3)}{\sum_w count(w_1, w_2, w)}$$

(the, green, *): 1,748× in EuroParl

w	count	p(w)
paper	801	0.458
group	640	0.367
light	110	0.063
party	27	0.015
ecu	21	0.012

LM QUALITY

We need to compare quality of various LMs.

2 approaches: extrinsic and intrinsic evaluation.

A good LM should assign a higher probability to a good (looking) text than to an incorrect text. For a fixed testing text we can compare various LMs.

ENTROPY

- Shannon, 1949
- the expected value (average) of the information contained in a message
- information viewed as the negative of the logarithm of the probability distribution
- events that always occur do not communicate information
- pure randomness has highest entropy (uniform distribution $log_2 n$)

$$E(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

PERPLEXITY

$$PP = 2^{H(p_{LM})}$$

$$PP(W) = p(w_1 w_2 \dots w_n)^{-\frac{1}{N}}$$

A good LM should not waste p for improbable phenomena. The lower entropy, the better $\longrightarrow$ the lower perplexity, the better.

Minimizing probabilities = minimizing perplexity.

WHAT INFLUENCES LM QUALITY?

- size of training data
- order of language model
- smoothing, interpolation, back-off

LARGE LM - N-GRAM COUNTS

How many unique n-grams are in a corpus?

order	types	singletons	%
unigram	86,700	33,447	(38,6%)
bigram	1,948,935	1,132,844	(58,1%)
trigram	8,092,798	6,022,286	(74,4%)
4-gram	15,303,847	13,081,621	(85,5%)
5-gram	19,882,175	18,324,577	(92,2%)

Taken from Europarl with 30 mil. tokens.

ZERO FREQUENCY, OOV, RARE WORDS

- probability must always be non zero
- to be able to measure perplexity
- maximum likelihood bad at it
- training data: work on Tuesday/Friday/Wednesday
- test data: work on Sunday,
  $p(Sunday|work\ on) = 0$

EXTRINSIC EVALUATION: SENTENCE COMPLETION

- Microsoft Research Sentence Completion Challenge
- evaluation of language models
- where perplexity not available
- from five Holmes novels
- training data: project Gutenberg

Model	**Acc**
Human	90
smoothed 3-gram	36
smoothed 4-gram	39
RNN	59
RMN (LSTM)	69

SENTENCE COMPLETION

The stage lost a fine XXX, even as science lost an acute reasoner, when he became a specialist in crime.
a) linguist b) hunter c) actor d) estate e) horseman

What passion of hatred can it be which leads a man to XXX in such a place at such a time.
a) lurk b) dine c) luxuriate d) grow e) wiggle

My heart is already XXX since i have confided my trouble to you.
a) falling b) distressed c) soaring d) lightened e) punished

My morning's work has not been XXX, since it has proved that he has the very strongest motives for standing in the way of anything of the sort.
a) invisible b) neglected c) overlooked d) wasted e) deliberate

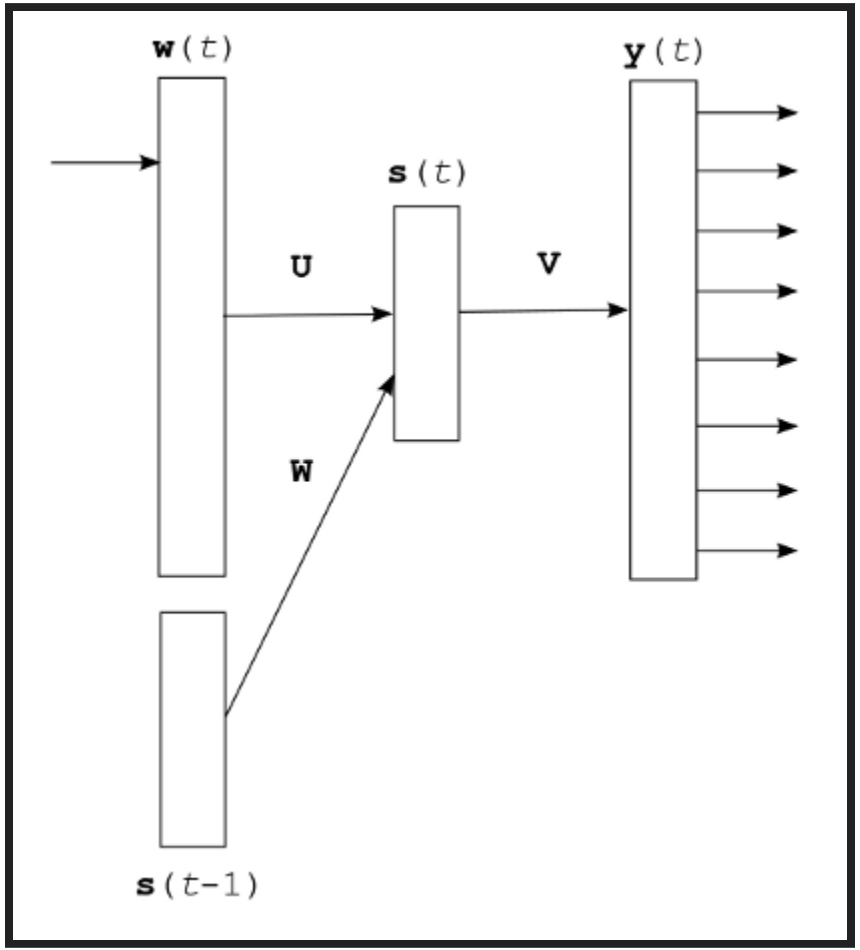
That is his XXX fault, but on the whole he's a good worker.
a) main b) successful c) mother's d) generous e) favourite

NEURAL NETWORK LANGUAGE MODELS

- old approach (1940s)
- only recently applied successfully to LM
- 2003 Bengio et al. (feed-forward NNLM)
- 2012 Mikolov (RNN)
- trending right now
- key concept: distributed representations of words
- 1-of-V, one-hot representation

RECURRENT NEURAL NETWORK

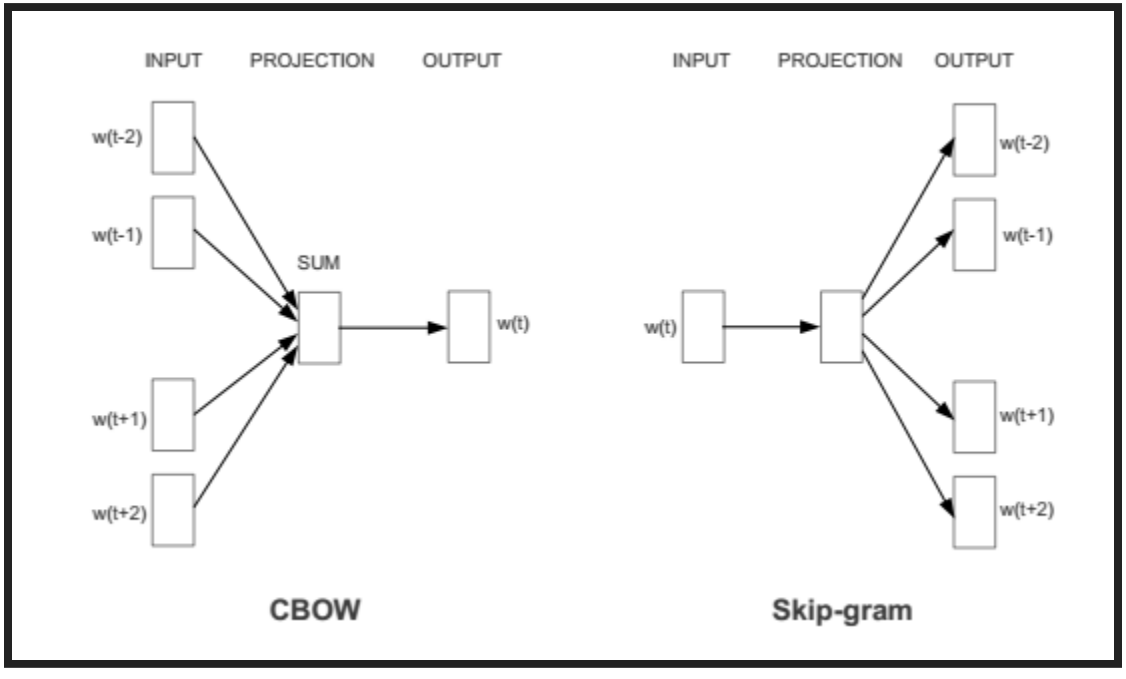
- Tomáš Mikolov (VUT)
- hidden layer feeds itself
- shown to beat n-grams by large margin

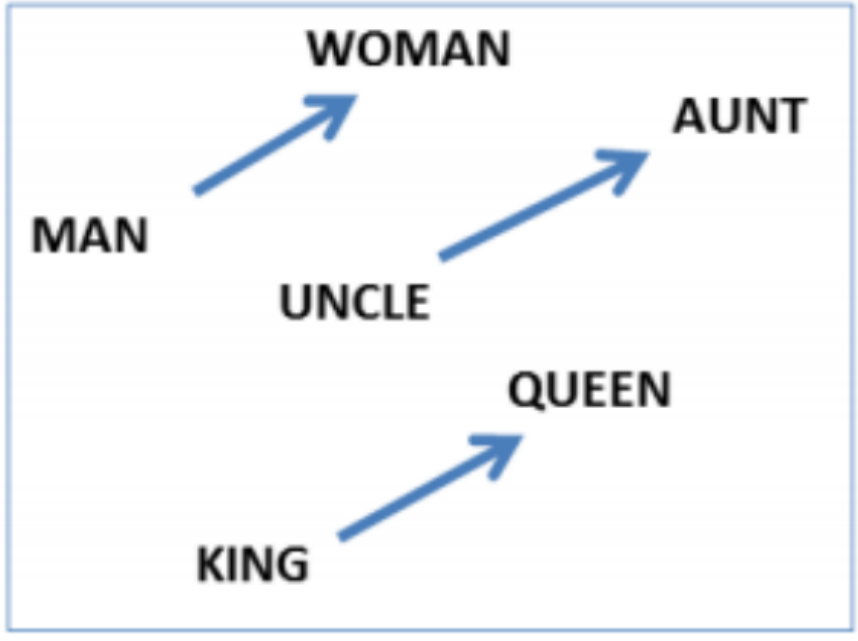


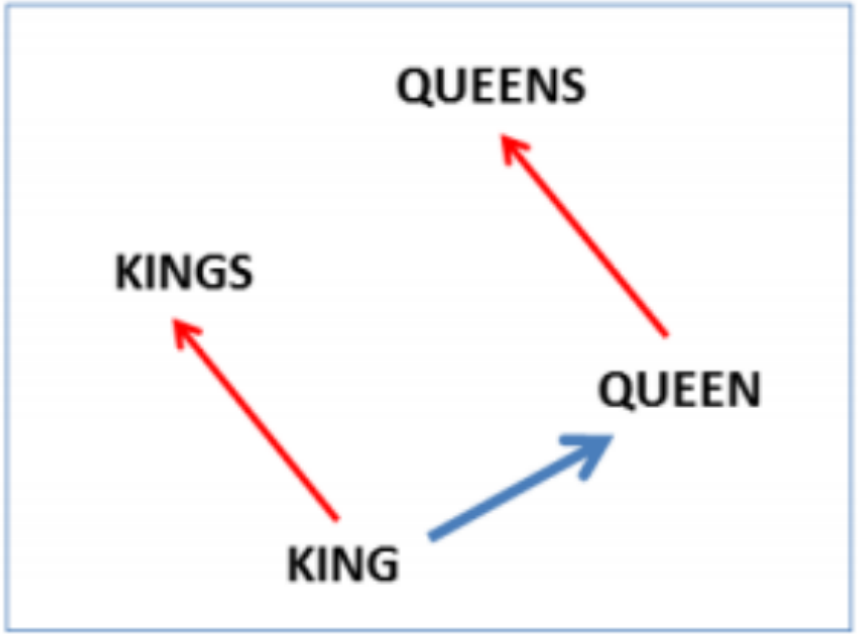
Model	Num. Params [billions]	Training Time [hours]	Training Time [CPUs]	Perplexity
Interpolated KN 5-gram, 1.1B n-grams (KN)	1.76	3	100	67.6
Katz 5-gram, 1.1B n-grams	1.74	2	100	79.9
Stupid Backoff 5-gram (SBO)	1.13	0.4	200	87.9
Interpolated KN 5-gram, 15M n-grams	0.03	3	100	243.2
Katz 5-gram, 15M n-grams	0.03	2	100	127.5
Binary MaxEnt 5-gram (n-gram features)	1.13	1	5000	115.4
Binary MaxEnt 5-gram (n-gram + skip-1 features)	1.8	1.25	5000	107.1
Hierarchical Softmax MaxEnt 4-gram (HME)	6	3	1	101.3
Recurrent NN-256 + MaxEnt 9-gram	20	60	24	58.3
Recurrent NN-512 + MaxEnt 9-gram	20	120	24	54.5
Recurrent NN-1024 + MaxEnt 9-gram	20	240	24	51.3

WORD EMBEDDINGS

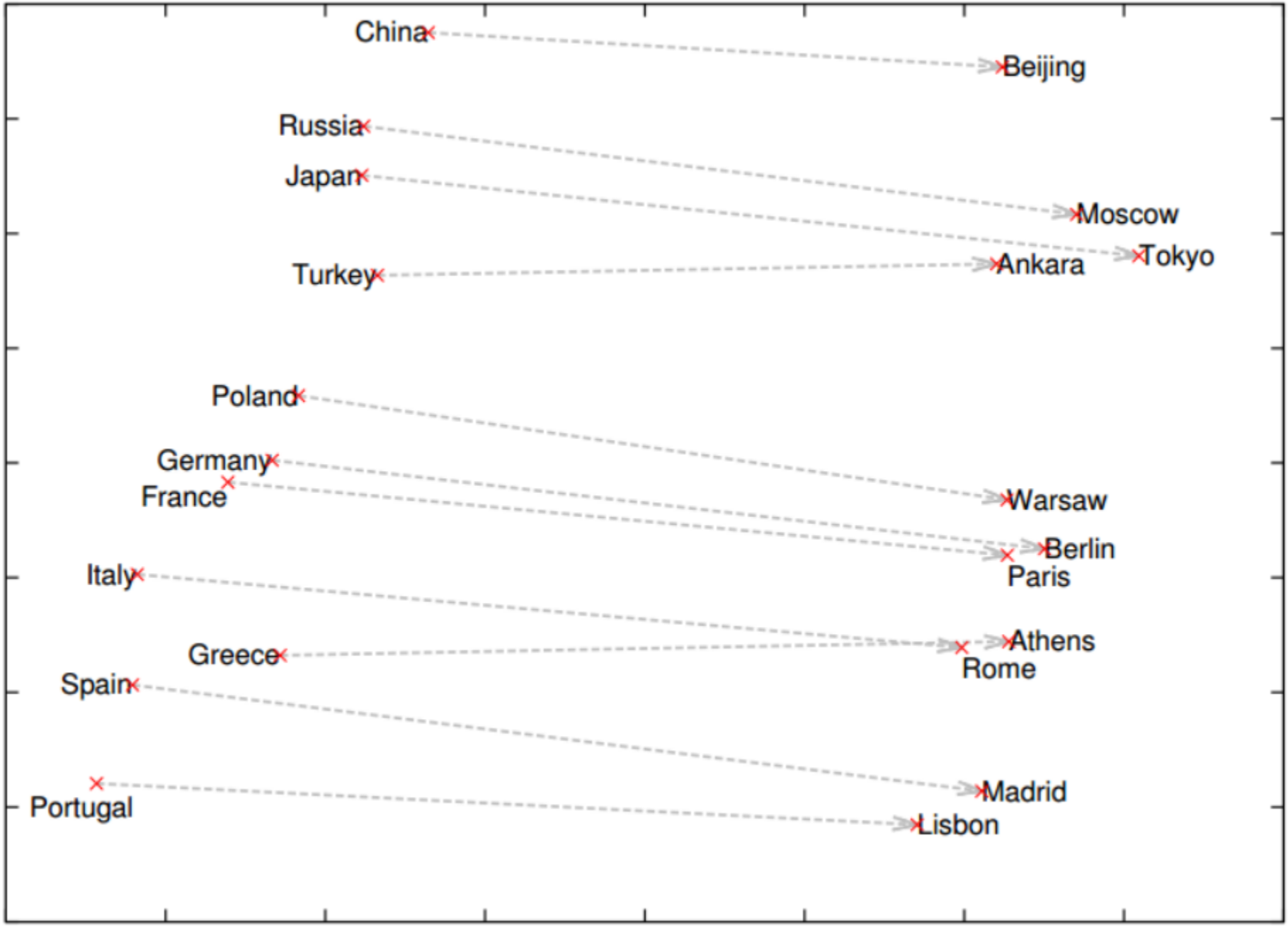
- distributional semantics with vectors
- skip-gram, CBOW (continuous bag-of-words)



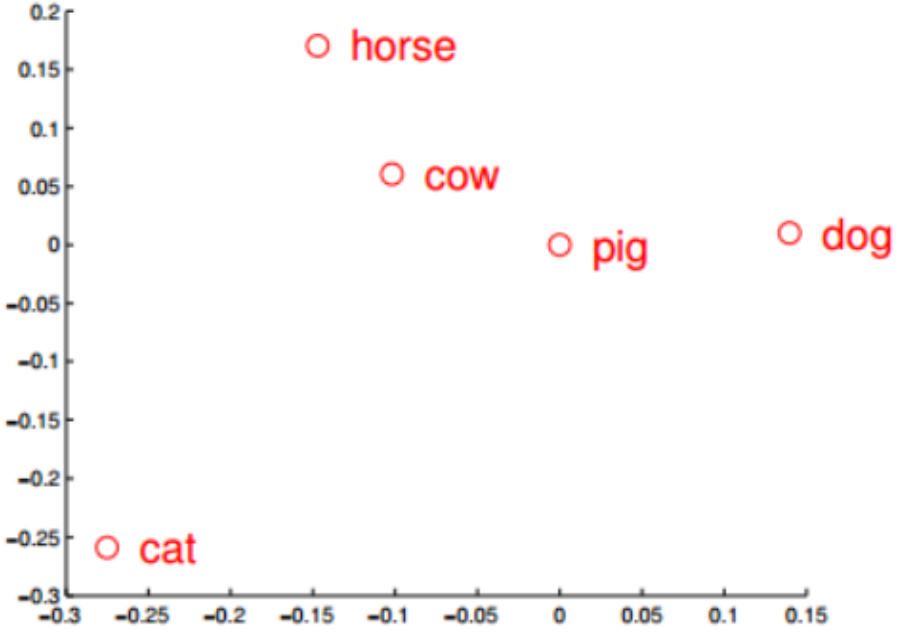

WOMAN
AUNT
MAN
UNCLE
QUEEN
KING

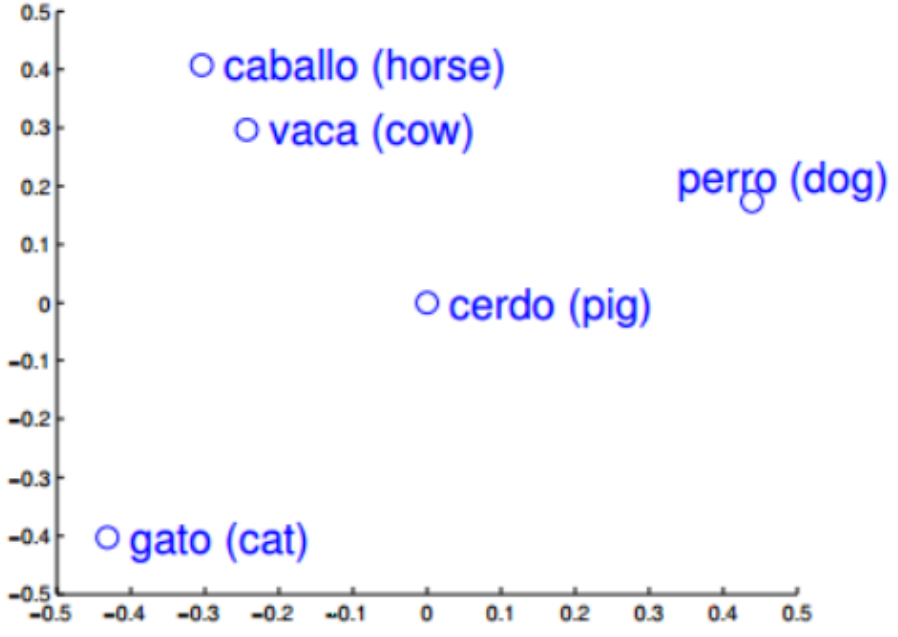

QUEENS
KINGS
QUEEN
KING

Expression	*Nearest token*
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs


China
Beijing
Russia
Moscow
Japan
Tokyo
Turkey
Ankara
Poland
Warsaw
Germany
Berlin
France
Paris
Italy
Athens
Greece
Rome
Spain
Madrid
Portugal
Lisbon

EMBEDDINGS IN MT





LONG SHORT-TERM MEMORY

- RNN model, can learn to memorize and learn to forget
- beats RNN in sequence learning
- LSTM

TRANSLATION MODELS

LEXICAL TRANSLATION

Standard lexicon does not contain information about frequency of translations of individual meanings of words.

key → klíč, tónina, klávesa

How often are the individual translations used in translations?

key → klíč (0.7), tónina (0.18), klávesa (0.11)

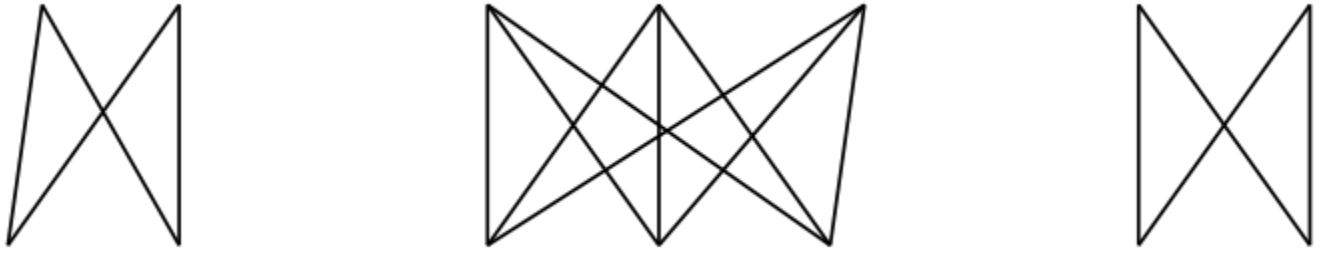
probability distribution p_f:

$$\sum_e p_f(e) = 1$$

$$\forall e : 0 \leq p_f(e) \leq 1$$

EM ALGORITHM - INITIALIZATION

... la maison ... la maison blue ... la fleur ...



... the house ... the blue house ... the flower ...

EM ALGORITHM - FINAL PHASE

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

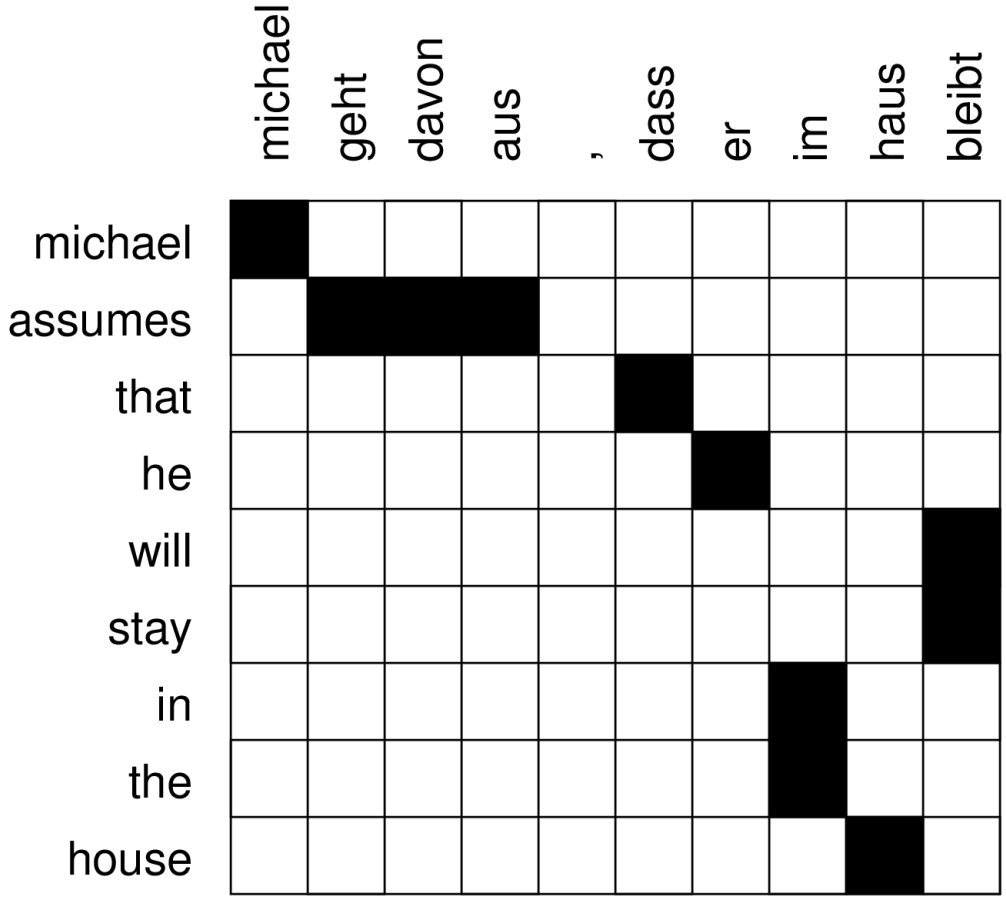
p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...

IBM MODELS

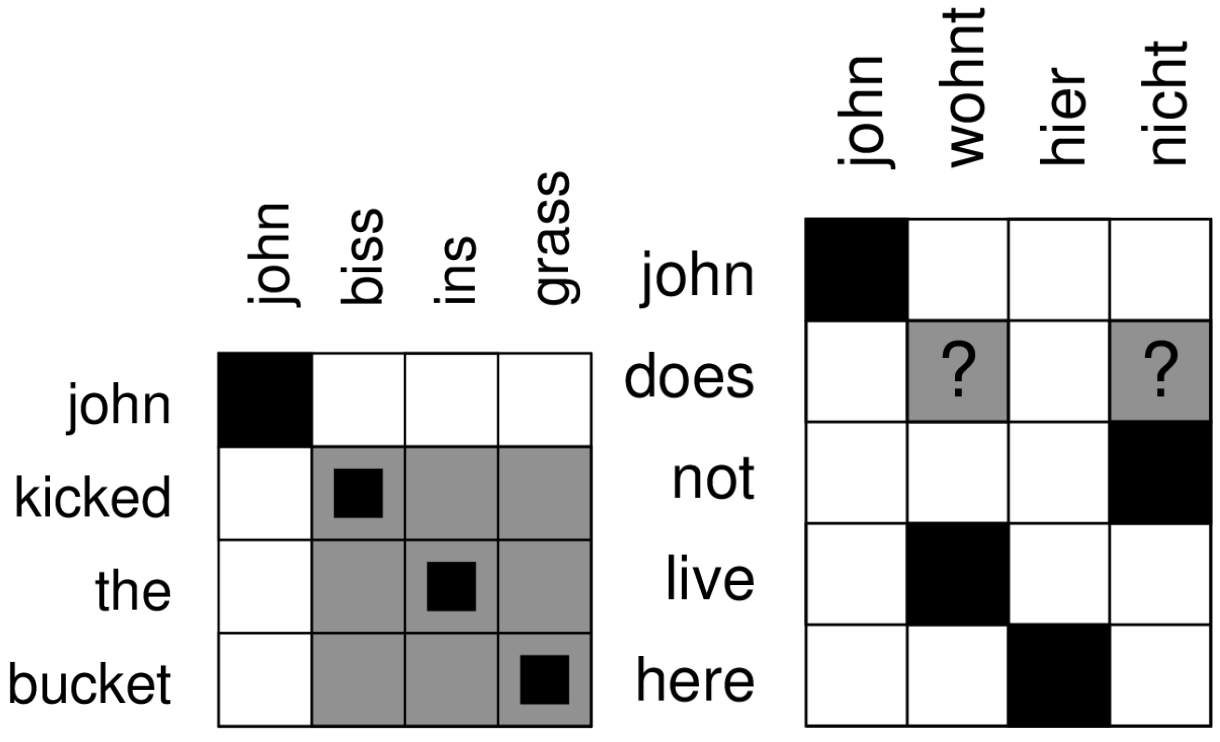
IBM-1 does not take context into account, cannot add and skip words. Each of the following models adds something more to the previous.

- IBM-1: lexical translation
- IBM-2: + absolute alignment model
- IBM-3: + *fertility* model
- IBM-4: + relative alignment model
- IBM-5: + further tuning

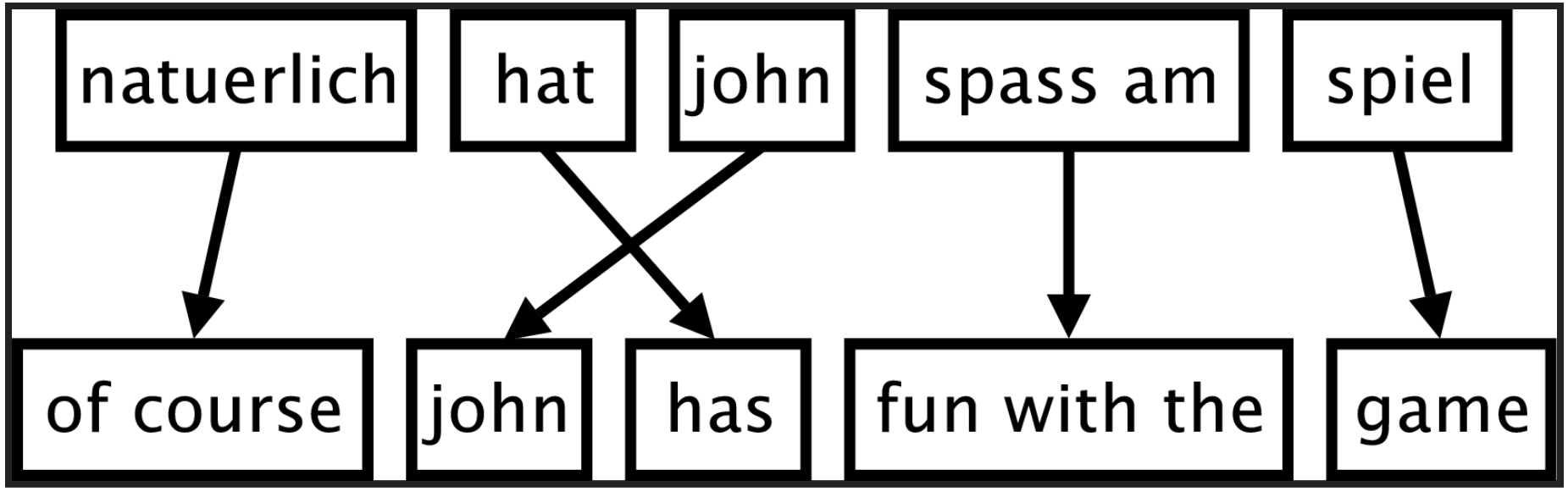
WORD ALIGNMENT MATRIX



WORD ALIGNMENT ISSUES



PHRASE-BASE TRANSLATION MODEL



Phrases not linguistically, but statistically motivated.
German *am* is seldom translated with single English *to*.
Cf. (fun (with (the game)))

ADVANTAGES OF PBTM

- translating n:m words
- word is not a suitable element for translation for many lang. pairs
- models learn to translate longer phrases
- simpler: no fertility, no NULL token etc.

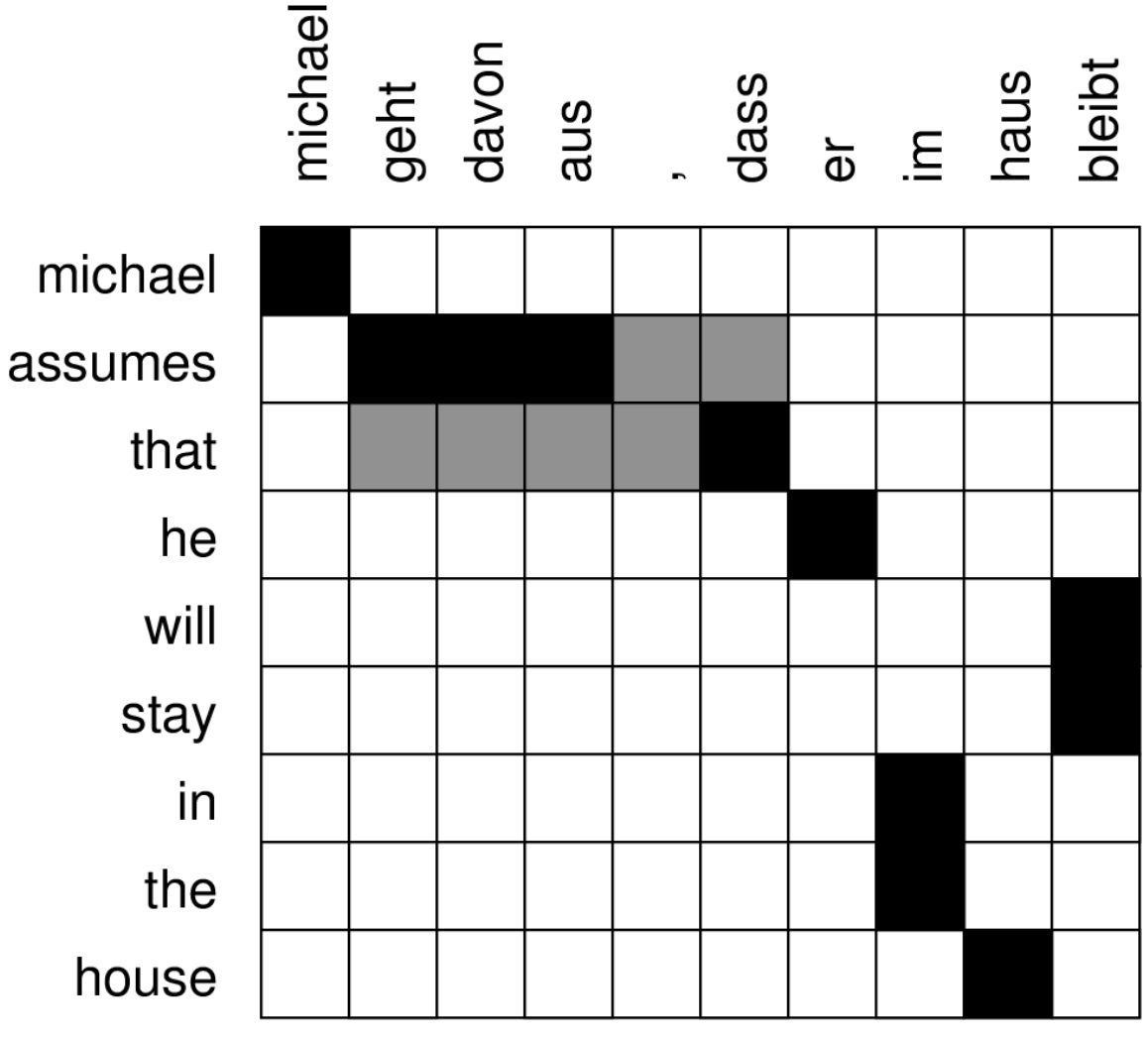
PHRASE-BASED MODEL

Translation probability $p(f|e)$ is split to phrases:

$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(start_i - end_{i-1} - 1)$$

Sentence f is split to I phrases $\bar{f}_i$, all segmentations are of the same probability. Function ϕ is translation probability for phrases. Function d is distance-based reordering model. $\mathbf{start}_i$ is position of the first word of phrase from sentence f, which is translated to i-th phrase of sentence e.

PHRASE EXTRACTION



EXTRACTED PHRASES

phr1	phr2
michael	michael
assumes	geht davon aus / geht davon aus
that	dass / , dass
he	er
will stay	bleibt
in the	im
house	haus
michael assumes	michael geht davon aus / michael geht davon aus ,
assumes that	geht davon aus , dass
assumes that he	geht davon aus , dass er
that he	dass er / , dass er

phr1	phr2
in the house	im haus
michael assumes that	michael geht davon aus , dass

PHRASE-BASED MODEL OF SMT

$$e^* = \mathbf{argmax}_e \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\, d(start_i - end_{i-1} - 1) \prod_{i=1}^{|\mathbf{e}|} p_{LM}(e_i|e_1 \ldots e_{i-1})$$

DECODING

Given a model p_{LM} and translation model $p(f|e)$ we need to find a translation with the highest probability but from exponential number of all possible translations.

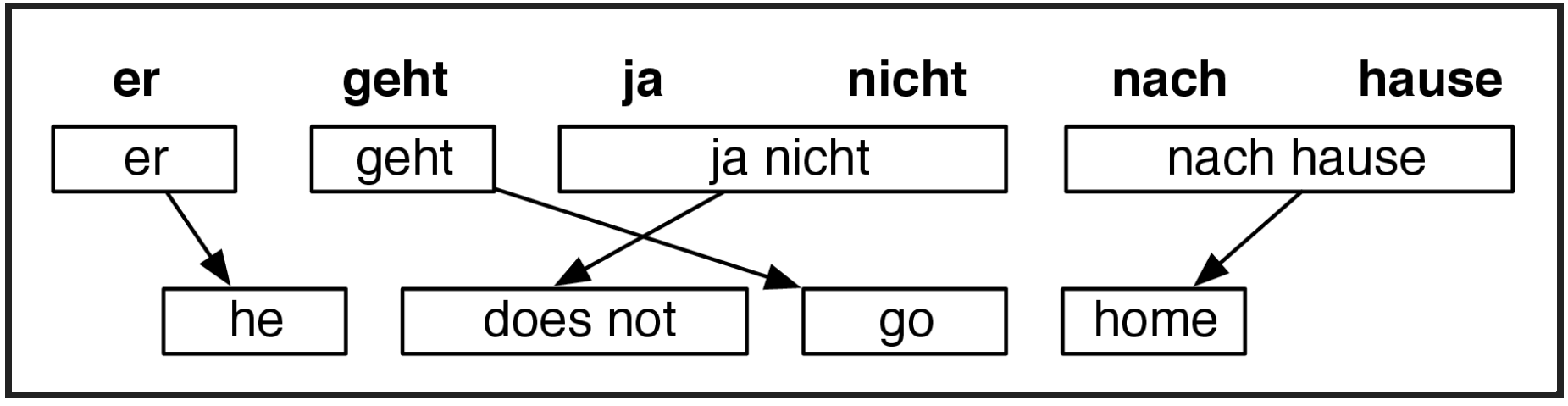
Heuristic search methods are used. It is not guaranteed to find the best translation.

Errors in translations are caused by
1) decoding process, when the best translation is not found owing to the heuristics or
2) models, where the best translation according to the probability functions is not the best possible.

EXAMPLE OF NOISE-INDUCED ERRORS (GOOGLE TRANSLATE)

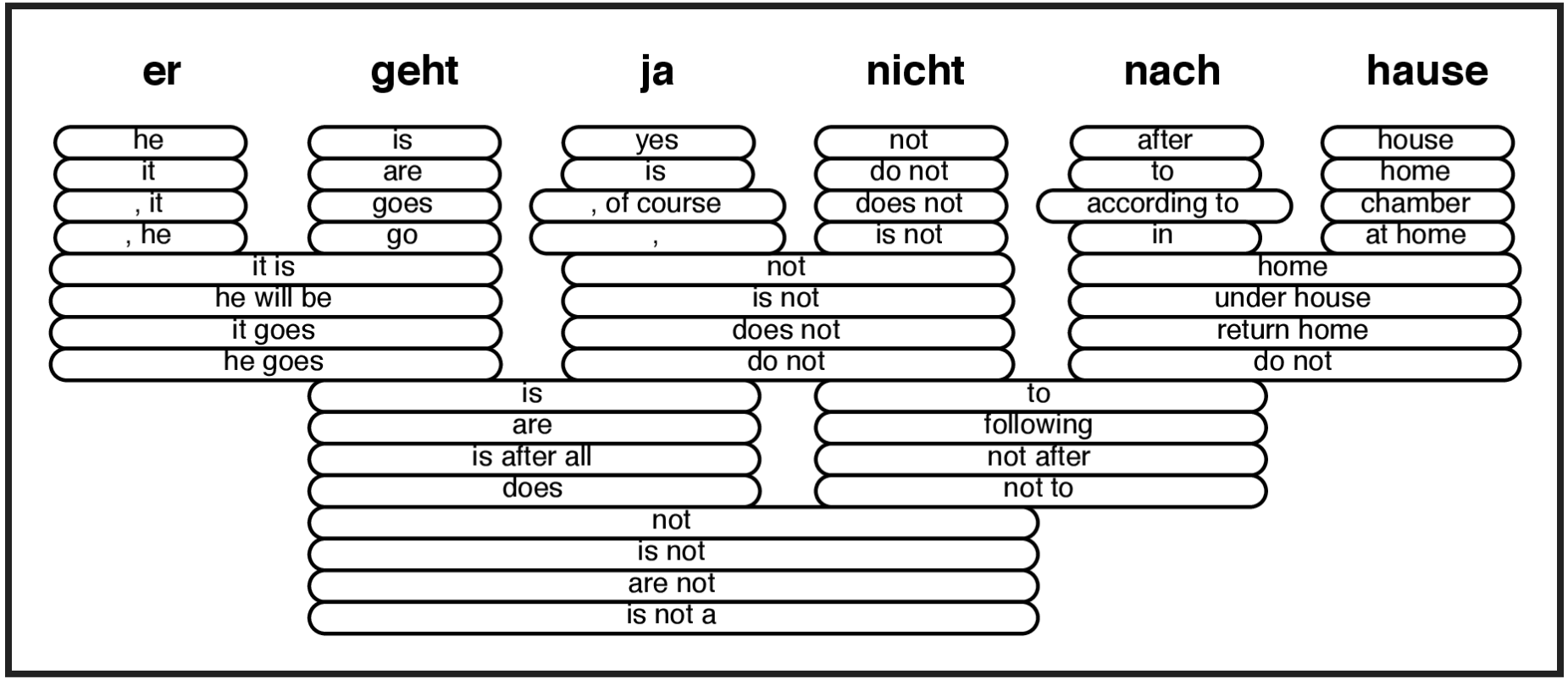
- Rinneadh clárúchán an úsáideora *yxc* a eiteach go rathúil.
- The user registration *yxc* made a successful rejection.
- Rinneadh clárúchán an úsáideora *qqq* a eiteach go rathúil.
- *Qqq* made registration a user successfully refused.

PHRASE-WISE SENTENCE TRANSLATION



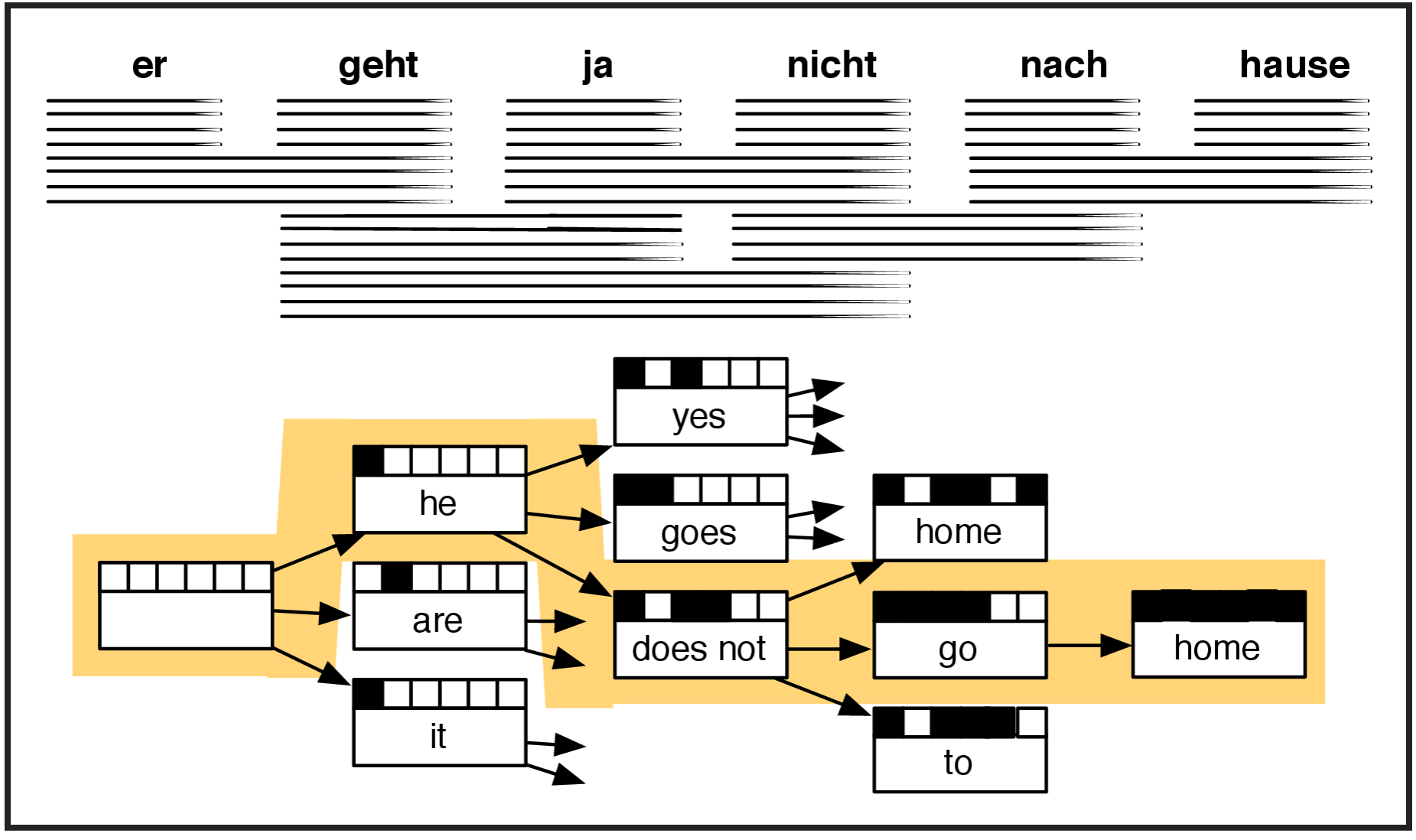
In each step of translation we count preliminary values of probabilities from the translation, reordering and language models.

SEARCH SPACE OF TRANSLATION HYPOTHESES



Exponential space of all possible translations → limit this space!

HYPOTHESIS CONSTRUCTION, BEAM SEARCH



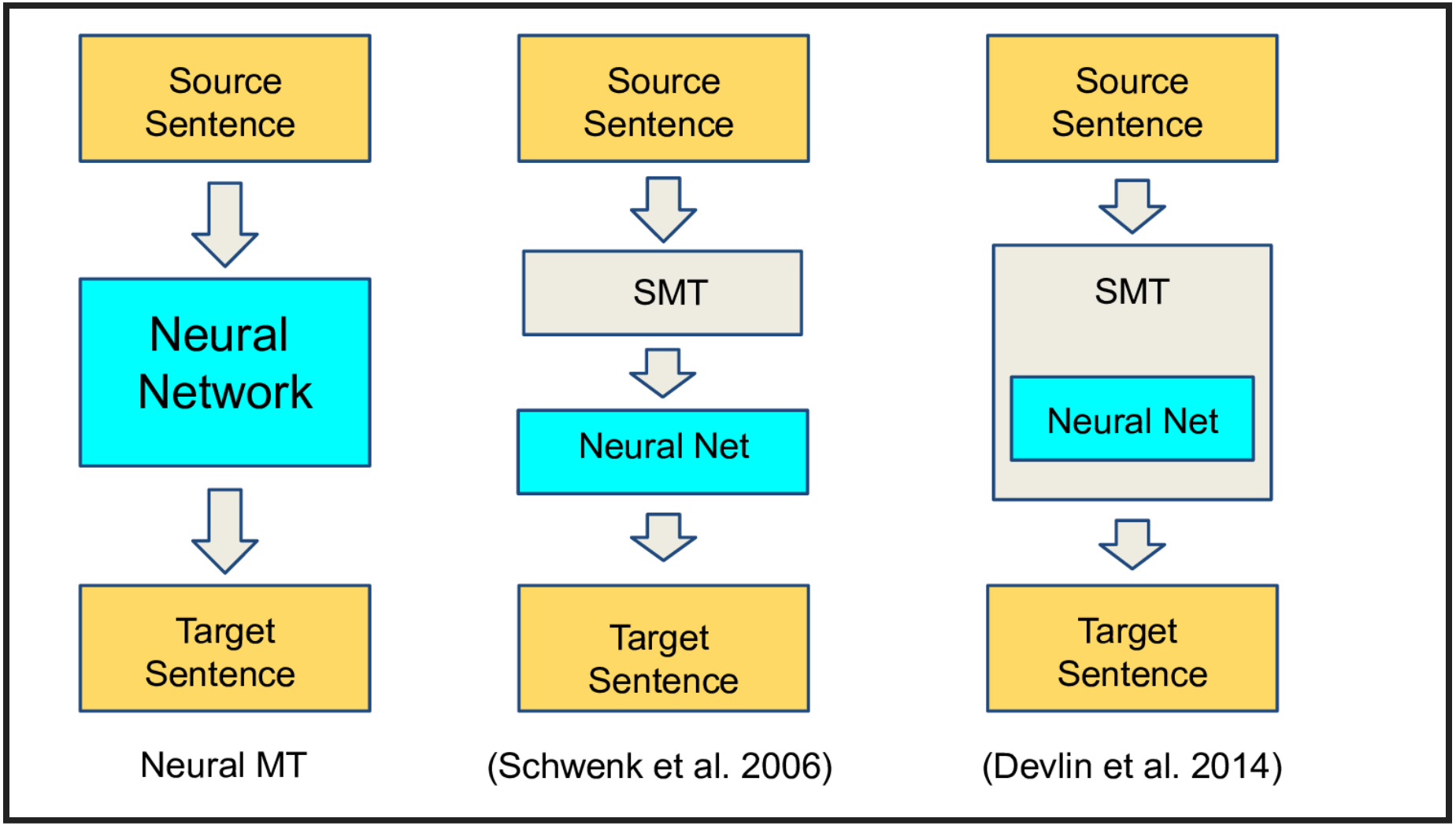
BEAM SEARCH

- *breadth-first* search
- on each level of the tree:
  generate all children of nodes on that level, sort them according to various heuristics
- store only a limited number of the best states on each level (beam width)
- only these states are investigated further
- the wider beam the smaller number of children are pruned
- with an unlimited width → breadth-first search
- the width correlates with memory consumption
- the best final state might not be found since it can be pruned

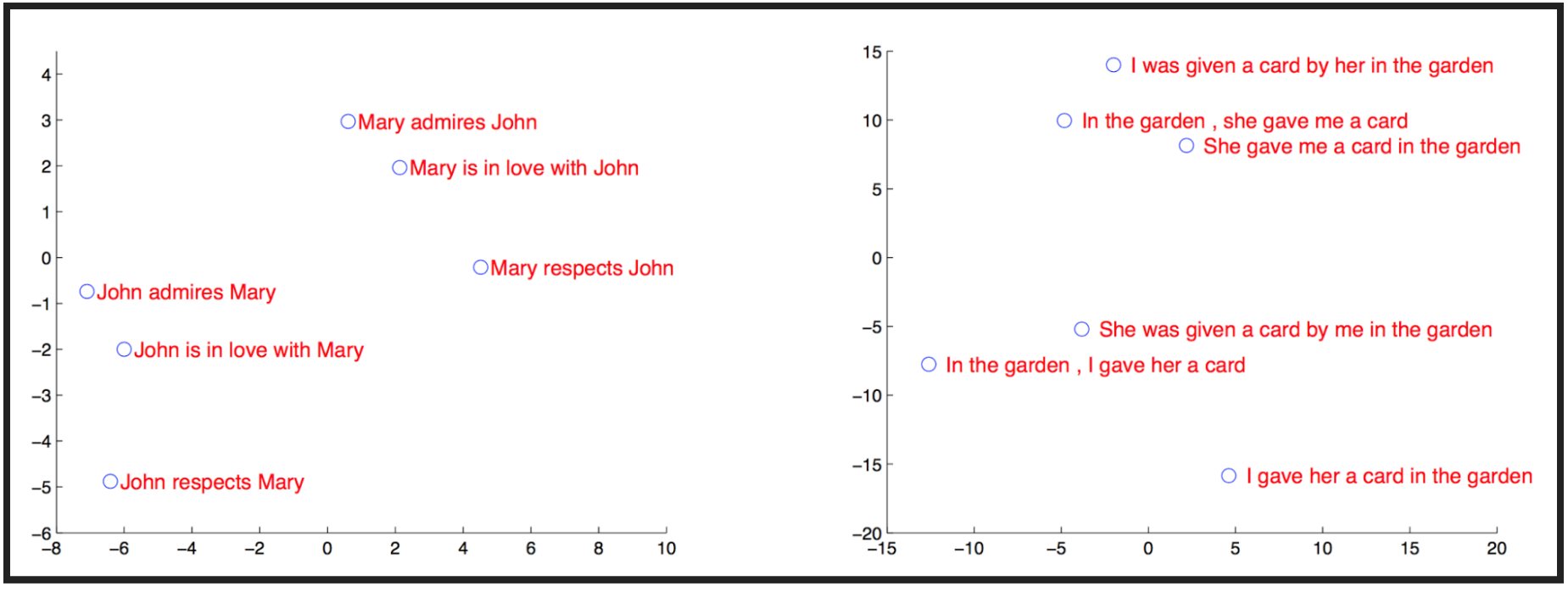
NEURAL NETWORK MACHINE TRANSLATION

- very close to state-of-the-art (PBSMT)
- a problem: variable length input and output
- learning to translate and align at the same time
- LISA
- hot topic (2014, 2015)

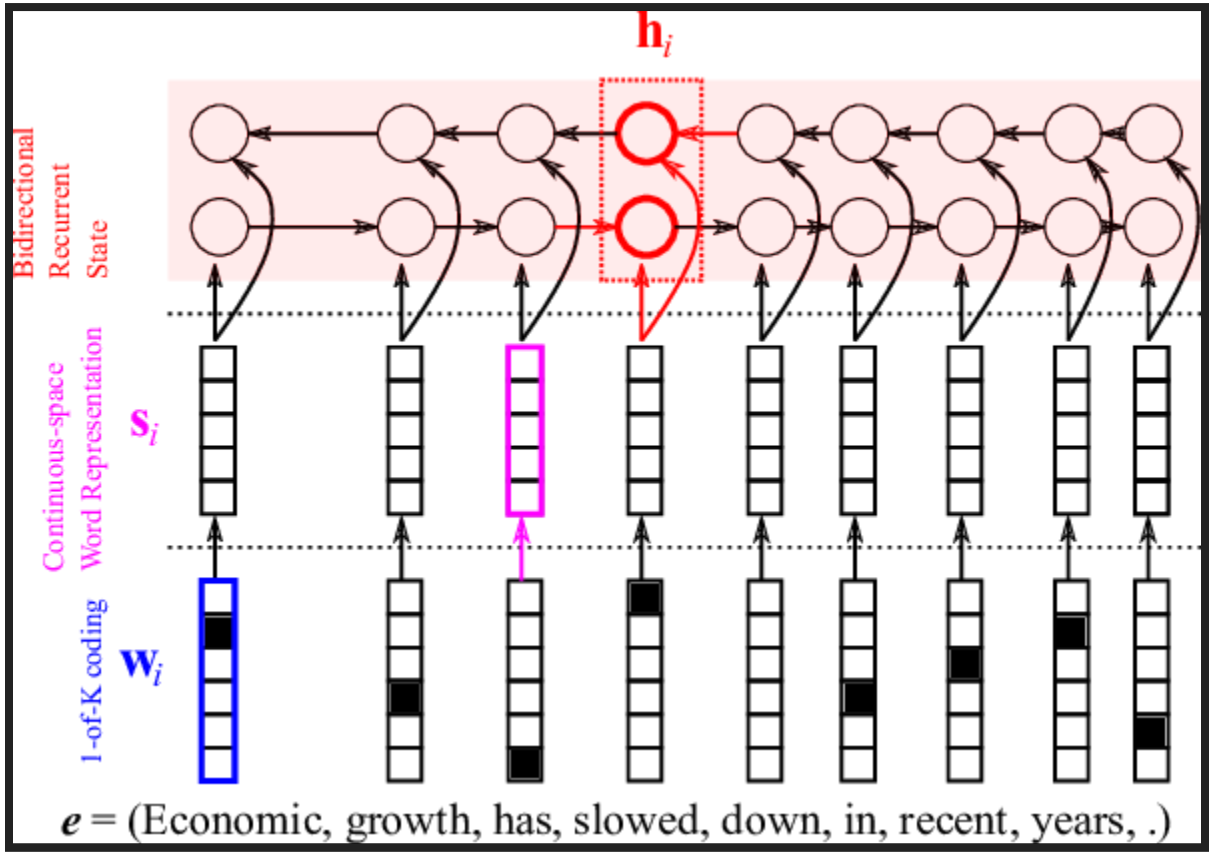
NN MODELS IN MT



SUMMARY VECTOR FOR SENTENCES

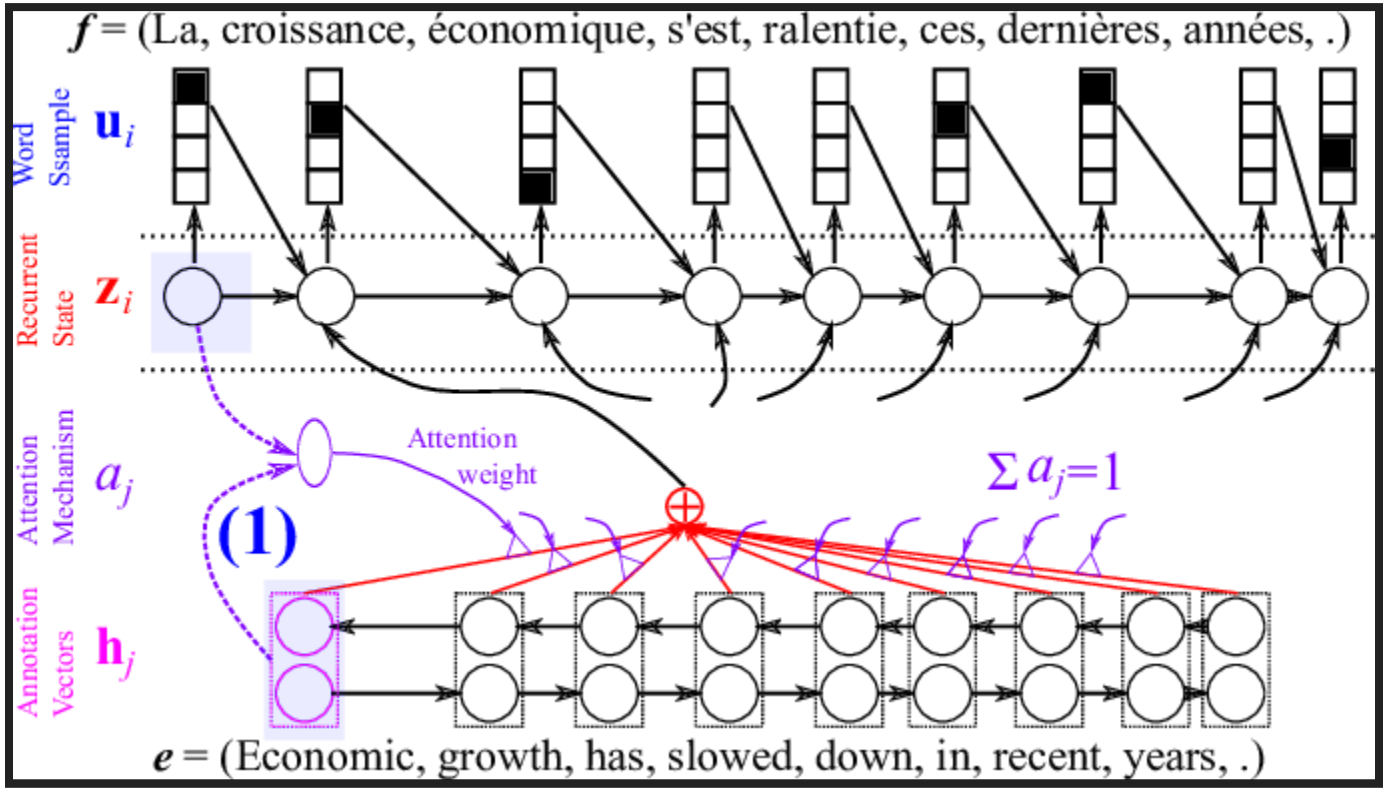


BIDIRECTIONAL RNN

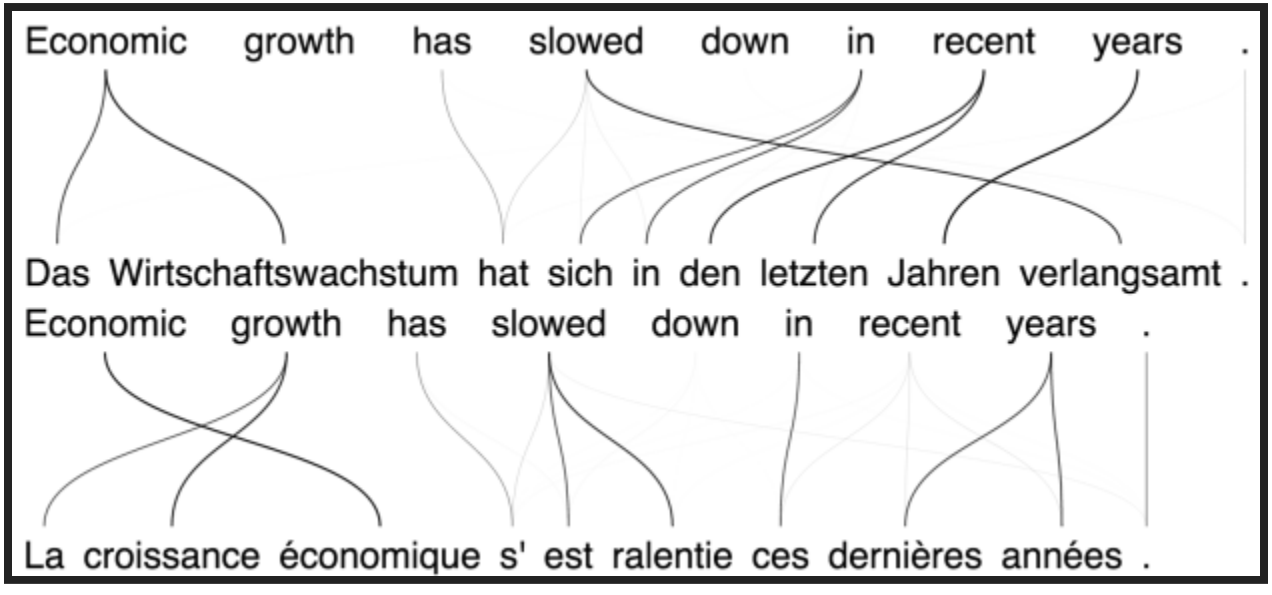


ATTENTION MECHANISM

A neural network with a single hidden layer, a single scalar output



ALIGNMENT FROM ATTETION



...more details here

ALIGNMENT WITH CO-OCCURRENCE STATISTICS

$$Dice = \frac{2f_{xy}}{f_x + f_y}$$

$$logDice = 14 + \log_2 D$$

biterms in SkE

MT QUALITY EVALUATION
OTHER MINOR TOPICS

MOTIVATION FOR MT EVALUATION

- **fluency**: is the translation fluent, in a natural word order?
- **adequacy**: does the translation preserve meaning?
- **intelligibility**: do we understand the translation?

EVALUATION SCALE

adequacy		**fluency**	
5	all meaning	5	flawless English
4	most meaning	4	good
3	much meaning	3	non-native
2	little meaning	2	dis-fluent
1	no meaning	1	incomprehensible

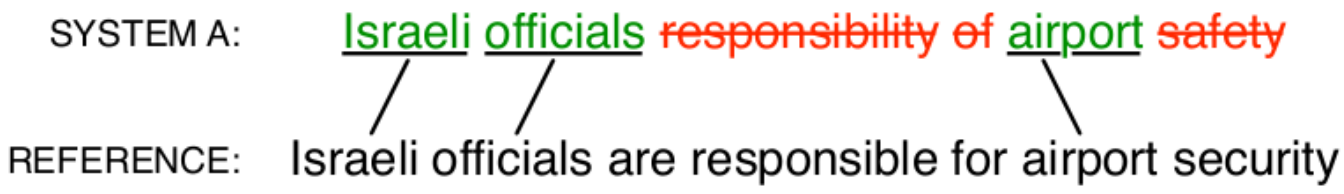
DISADVANTAGES OF MANUAL EVALUATION

- slow, expensive, subjective
- inter-annotator agreement (IAA) shows people agree more on fluency than on adequacy
- another option: is X better than Y? → higher IAA
- **or** time spent on post-editing
- **or** how much cost of translation is reduced

AUTOMATIC TRANSLATION EVALUATION

- advantages: speed, cost
- disadvantages: do we really measure quality of translation?
- gold standard: manually prepared reference translations
- candidate $\boldsymbol{c}$ is compared with $\boldsymbol{n}$ reference translations $\boldsymbol{r_i}$
- the paradox of automatic evaluation: the task corresponds to situation where students are to assess their own exam: how they know where they made a mistake?
- various approaches: n-gram shared between $\boldsymbol{c}$ and $\boldsymbol{r_i}$, edit distance, ...

RECALL AND PRECISION ON WORDS



$$\text{precision} = \frac{\text{correct}}{\text{output-length}} = \frac{3}{6} = 50\%$$

$$\text{recall} = \frac{\text{correct}}{\text{reference-length}} = \frac{3}{7} = 43\%$$

$$\text{f-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2 \times \frac{.5 \times .43}{.5 + .43} = 46\%$$

RECALL AND PRECISION: SHORTCOMINGS

metrics	system A	system B
precision	50%	100%
recall	43%	100%
f-score	46%	100%

It does not capture wrong word order.

BLEU

- standard metrics (2001)
- IBM, Papineni
- n-gram match between reference and candidate translations
- precision is calculated for 1-, 2- ,3- and 4-grams
- + **brevity penalty**

$$\mathrm{BLEU} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right) \left(\prod_{i=1}^{4} \text{precisio}\right.$$

BLEU: AN EXAMPLE

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (4-GRAM MATCH)

metrics	system A	system B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

NIST

- NIST: National Institute of Standards and Technology
- weighted matches of n-grams (information value)
- very similar results as for BLEU (a variant)

NEVA

- Ngram EVAluation
- BLEU score adapted for short sentences
- it takes into account synonyms (stylistic richness)

WAFT

- Word Accuracy for Translation
- edit distance between $\boldsymbol{c}$ and $\boldsymbol{r}$

$$\mathbf{WAFT} = 1 - \frac{d+s+i}{max(l_r,l_c)}$$

TER

- Translation Edit Rate
- the least edit steps (deletion, insertion, swap, replacement)
- $\boldsymbol{r} =$ dnes jsem si při fotbalu zlomil kotník
- $\boldsymbol{c} =$ při fotbalu jsem si dnes zlomil kotník
- TER = ?

$$\mathbf{TER} = \frac{\mathbf{number\ of\ edits}}{\mathbf{avg.\ number\ of\ ref.\ words}}$$

HTER

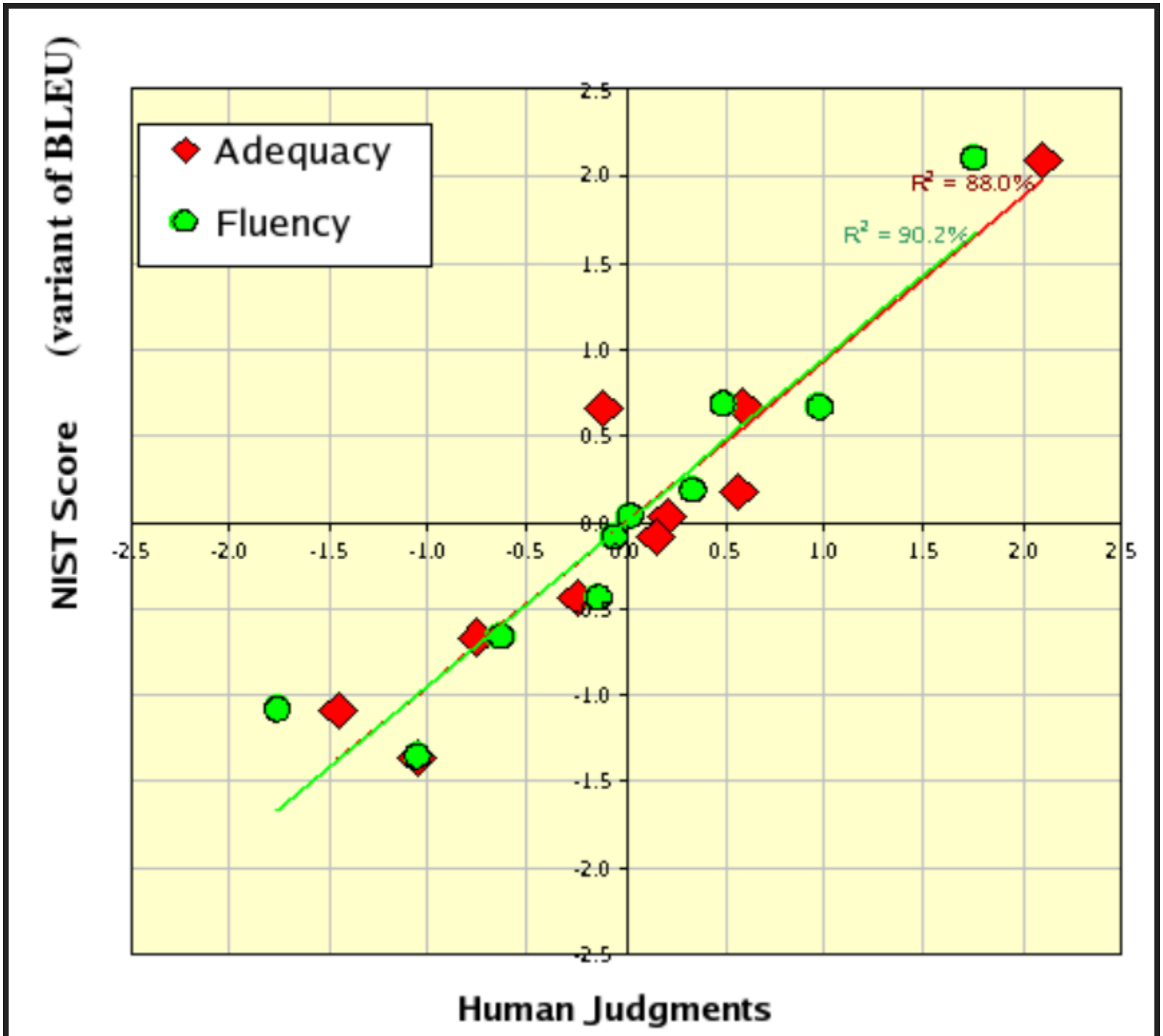
- Human TER
- r manually prepared and then TER is applied

METEOR

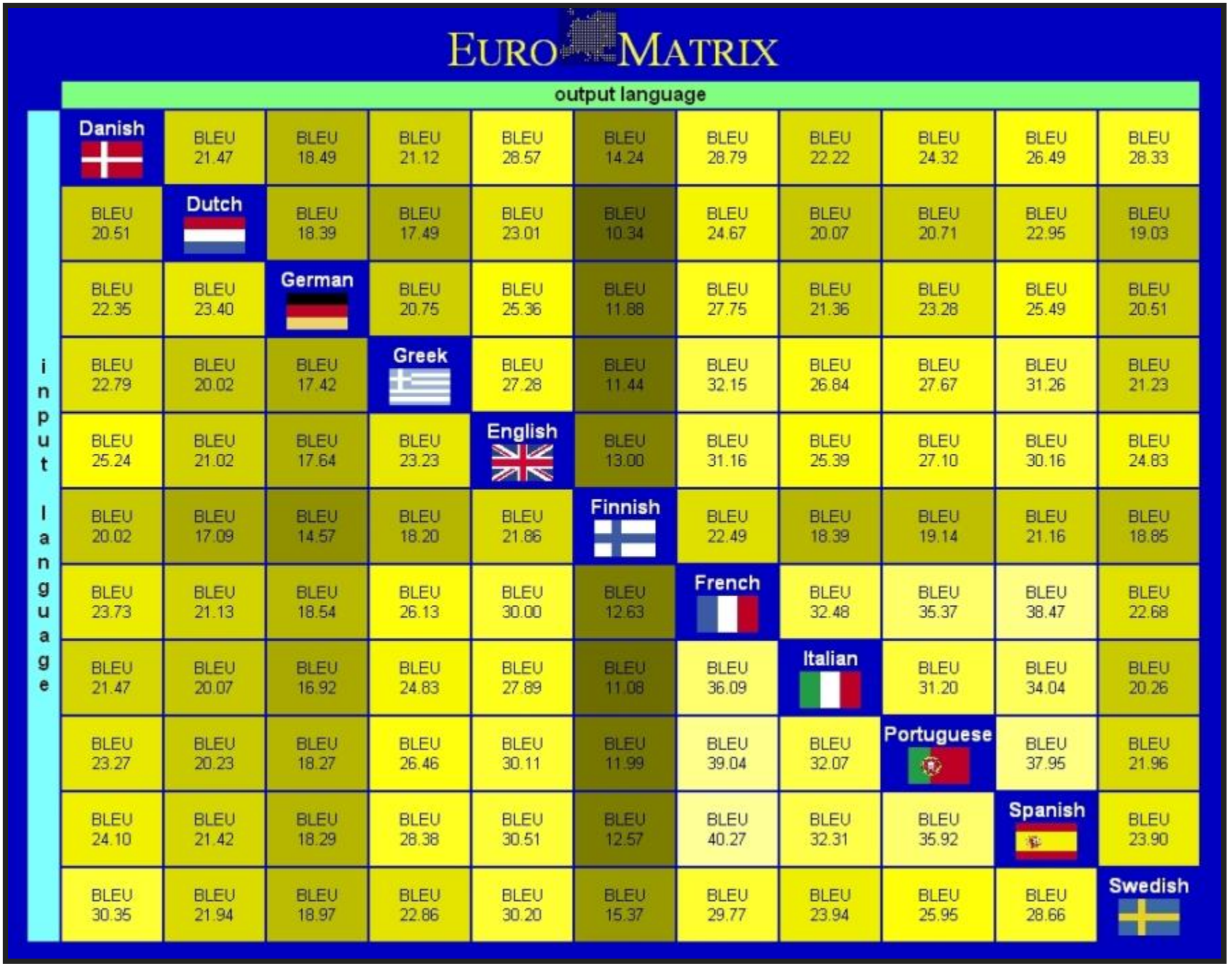
- aligns hypotheses to one or more references
- exact, stem (morphology), synonym (WordNet), paraphrase matches
- various scores including WMT ranking and NIST adequacy
- extended support for English, Czech, German, French, Spanish, and Arabic.
- high correlation with human judgments

EVALUATION OF EVALUATION METRICS

Correlation of automatic evaluation with manual evaluation.


NIST Score (variant of BLEU)
Adequacy
Fluency
R² = 88.0%
R² = 90.2%
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Human Judgments

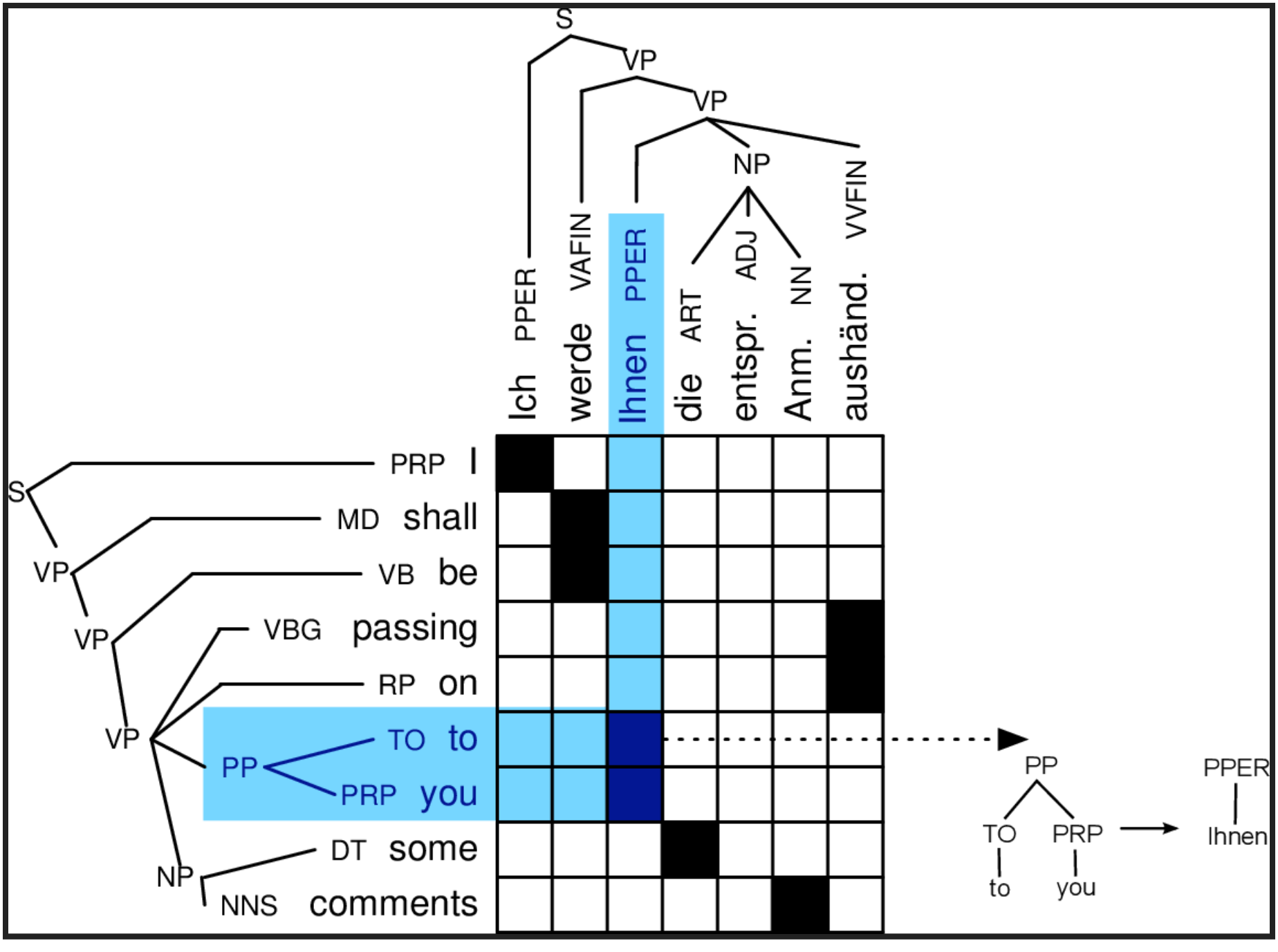
EUROMATRIX


EURO MATRIX
output language
input language
Danish
Dutch
German
Greek
English
Finnish
French
Italian
Portuguese
Spanish
Swedish
BLEU 21.47 BLEU 18.49 BLEU 21.12 BLEU 28.57 BLEU 14.24 BLEU 28.79 BLEU 22.22 BLEU 24.32 BLEU 26.49 BLEU 28.33
BLEU 20.51 BLEU 18.39 BLEU 17.49 BLEU 23.01 BLEU 10.34 BLEU 24.67 BLEU 20.07 BLEU 20.71 BLEU 22.95 BLEU 19.03
BLEU 22.35 BLEU 23.40 BLEU 20.75 BLEU 25.36 BLEU 11.88 BLEU 27.75 BLEU 21.36 BLEU 23.28 BLEU 25.49 BLEU 20.51
BLEU 22.79 BLEU 20.02 BLEU 17.42 BLEU 27.28 BLEU 11.44 BLEU 32.15 BLEU 26.84 BLEU 27.67 BLEU 31.26 BLEU 21.23
BLEU 25.24 BLEU 21.02 BLEU 17.64 BLEU 23.23 BLEU 13.00 BLEU 31.16 BLEU 25.39 BLEU 27.10 BLEU 30.16 BLEU 24.83
BLEU 20.02 BLEU 17.09 BLEU 14.57 BLEU 18.20 BLEU 21.86 BLEU 22.49 BLEU 18.39 BLEU 19.14 BLEU 21.16 BLEU 18.85
BLEU 23.73 BLEU 21.13 BLEU 18.54 BLEU 26.13 BLEU 30.00 BLEU 12.63 BLEU 32.48 BLEU 35.37 BLEU 38.47 BLEU 22.68
BLEU 21.47 BLEU 20.07 BLEU 16.92 BLEU 24.83 BLEU 27.89 BLEU 11.08 BLEU 36.09 BLEU 31.20 BLEU 34.04 BLEU 20.26
BLEU 23.27 BLEU 20.23 BLEU 18.27 BLEU 26.46 BLEU 30.11 BLEU 11.99 BLEU 39.04 BLEU 32.07 BLEU 37.95 BLEU 21.96
BLEU 24.10 BLEU 21.42 BLEU 18.29 BLEU 28.38 BLEU 30.51 BLEU 12.57 BLEU 40.27 BLEU 32.31 BLEU 35.92 BLEU 23.90
BLEU 30.35 BLEU 21.94 BLEU 18.97 BLEU 22.86 BLEU 30.20 BLEU 15.37 BLEU 29.77 BLEU 23.94 BLEU 25.95 BLEU 28.66

EUROMATRIX II

	Target language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN		40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3		38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3		35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6		43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7		34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5		54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4		25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4		37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9		29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0		25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7		33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2		29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8		40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	25.3	34.2	32.4	35.6	29.3	38.9	38.4		23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2		44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0		32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8		44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5		39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5		31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3		42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9		42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	

SYNTACTIC RULES EXTRACTION


S
VP
VP
NP
Ich PPER
werde VAFIN
Ihnen PPER
die ART
entspr. ADJ
Anm. NN
aushänd. VVFIN
S
VP
VP
VP
PP
NP
PRP I
MD shall
VB be
VBG passing
RP on
TO to
PRP you
DT some
NNS comments
PP
TO
PRP
to
you
PPER
Ihnen

HYBRID SMT+RBMT

- Chimera, UFAL
- TectoMT + Moses
- better than Google Translate (En-Cz)